ONLINE SOCIAL NETWORKS: REAL-TIME GRAPH SAMPLING, MULTILAYERED INFORMATION DIFFUSION, AND EQUALITY ISSUES

Ph.D. Dissertation

Giannis Haralabopoulos

Presented to the

Department of Computer Science and Biomedical Informatics

University of Thessaly

2016

ONLINE SOCIAL NETWORKS: REAL-TIME GRAPH SAMPLING, MULTILAYERED INFORMATION DIFFUSION, AND EQUALITY ISSUES

Ph.D. Dissertation

Giannis Haralabopoulos

Presented to the

Department of Computer Science and Biomedical Informatics

University of Thessaly

2016

University of Thessaly

2016

Dedication

To my family, friends and colleagues.

Acknowledgements

I would like to thank my supervisor, Prof. Ioannis Anagnostopoulos for his confidence, encouragement, expertise, advices and support, throughout the duration of my Ph.D. studies.

ONLINE SOCIAL NETWORKS: REAL-TIME GRAPH SAMPLING, MULTILAYERED INFORMATION DIFFUSION, AND EQUALITY ISSUES

Committee:

Ioannis Anagnostopoulos
Assist. Professor
University of Thessaly
(Supervisor)

Christos Douligeris
Professor
University of Piraeus

Angelos Michalas
Assoc. Professor
Technological Educational
Institute of Western Macedonia

Online Social Networks: Real-time Graph Sampling, Multilayered Information Diffusion, and Equality Issues

Giannis Haralabopoulos
University of Thessaly, 2016
Supervisor: Ioannis Anagnostopoulos

Over the last decade, Online Social Networks (OSNs) evolved into a new medium of communication, social connection and information sharing. Our research, aims to study the structure, properties, information diffusion, bias and features of these networks. The dissertation is split in three discrete sections, with each one covering a specific research topic, corresponding to our scientific contributions. Which are in; data sampling methods for online restrictive environments, inter-OSNs information propagation, OSN issues and future possibilities. The first section covers the issues of OSN real time sampling. Prior to the introduction of a novel sampling method, an extensive literature review is presented. In order to test our proposal, we used multiple up-to-date Twitter samples to compare it with established sampling methods. Our second research subject covers information diffusion and sharing, in multiple OSNs. At the time, no published research has considered information diffusion over multiple online networks. In addition to the literature overview, we also monitored and studied diverse topics in three OSNs for a period of sixty days. The respective results are also detailed in this section. Lastly, throughout our extensive engagement with modern OSNs, we observed multiple issues and shortcomings (such as privacy, bias and data access) prevalent in most of them. These issues, along with the concept of a new anonymous OSN, are presented in the third research topic. Our contributions provide deep insight and solutions in the broader field of OSNs, with subjects ranging from data mining and information diffusion, to social bias and online equality.

Table of Contents

List of Tables 14
List of Figures 15
1. INTRODUCTION 16
2. SOCIAL NETWORK SAMPLING 18
2.1. Online social networks 18
2.1.1. Introduction 18
2.1.2. The History Behind Social Networking 19
2.1.3. The Internet History 32
2.1.3.1. Early Forms of Online Social Networks 36
2.1.3.2. Modern Online Social Networks 38
2.1.3.2.1. MySpace 40
2.1.3.2.2. LinkedIn 41
2.1.3.2.3. Facebook 41
2.1.3.2.4. Microsoft and Yahoo 42
2.1.3.2.5. Tencent Qzone 43
2.1.3.2.6. Flickr and Youtube 43
2.1.3.2.7. Reddit 45
2.1.3.2.8. Twitter 45
2.1.3.2.9. Google Plus 46
2.1.3.3. The present state of Online Social Networking 47
2.2. Online Social networks as graphs 50
2.2.1. Graph Theory 50
2.2.1.1. Graph Drawing 54
2.2.2. Graph sampling 57
2.2.3. Graph analysis 59
2.2.3.1. Related Work 60
2.3. Real Time Enhanced Random Sampling 64
2.3.1. Problems of OSN analysis and sampling 64
2.3.2. Proposed Enhancements 65
2.3.2.1.1. Selective Sampling 66
2.3.2.1.2. Enhanced Subgraph 67
2.3.3. The algorithm 68
2.3.3.1. Description 69
2.3.4. Dataset 73
2.3.5. Analysis tools and evaluation 74

2.3.6. Experimentation 76
2.3.7. Results 78
2.3.7.1. RNS with Enhanced Subgraph Enhancement 81
2.3.7.2. RNS with Selective Sampling Enhancement 84
2.3.7.3. Stanford Large Network Dataset Collection 92
2.3.8. Conclusions 97
3. MULTI LAYERED INFORMATION DIFFUSION 99
3.1. Information networks 99
3.1.1. Related Work 100
3.2. Networks of Interest 106
3.3. Lifespan and propagation of Information in Online Social Networks: A Case Study based on Reddit 109
3.3.1. Methodology 109
3.3.1.1. Units of Interest 112
3.3.1.2. The Virality Criterion 113
3.3.2. The monitoring algorithm 114
3.3.3. Results 115
3.3.3.1. Topics 115
3.3.3.2. OSN Diffusion 120
3.3.4. Diffusion over time 126
3.3.5. Diffusion Patterns 128
3.3.5.1. Pattern 1 129
3.3.5.2. Pattern 2 132
3.3.5.3. Pattern 3 134
3.3.5.4. Pattern 4 136
3.3.5.5. Pattern 5 138
3.3.6. Information Diffusion 140
3.3.6.1. The ImgUr Diffusion Model 141
3.3.6.2. The YouTube Diffusion Model 144
3.3.6.3. Extended UoI Measurement 146
3.3.7. Discussion 150
4. AN ANONYMOUS ONLINE SOCIAL NETWORK OF OPINIONS 154
4.1. Online Social Parity 154
4.1.1. Social Issues 154
4.1.2. Online Social Issues 161
4.1.2.1. Privacy 161

4.1.2.2. Tightly Knit Communities 162
4.1.2.3. Power Law Distribution 163
4.1.2.4. Bias 165
4.1.2.5. Credibility 165
4.1.2.6. Information Quality 166
4.1.2.7. Various Issues 167
4.2. The Online Social Experiment 169
4.2.1. Equality, Privacy and Data 169
4.2.2. Design 169
4.2.3. Aim and Necessity 171
5. CONTRIBUTIONS 172
5.1. Contributions 172
5.2. Future Research 173
BIBLIOGRAPHY 177

List of Tables

Table 1. Test graphs mean and marginal values 76
Table 2. Every iteration of RNS with 10% sample size in the first TG 78
Table 3. Random Node Sampling results for the first TG 79
Table 4. Percentage error of RNS in the first TG 80
Table 5. Mean percentage error 81
Table 6. Mean percentage error of RNS 85-15 84
Table 7. Mean percentage error of RNS 80-20 and RNS 75-25 87
Table 8. RNS and RNS 80-20 comparison, with or without Enhanced subgraph enhancement 90
Table 9. Mean inaccuracy of Selective Sampling RNS 91
Table 10. Number and percentage of less vertices used in Selective Sampling 92
Table 11. Detailed properties of every SG 93
Table 12. Estimated and actual distribution of every SG 94
Table 13. Mean percentage error for multi component graphs 95
Table 14. Mean percentage error for single component graphs 96
Table 15. Posts of Interest Domain Distribution 116
Table 16. Posts sorted by Topic, Category and Domain 118
Table 17. Post mentions per Topic 121
Table 18. "New" category diffusion percentages 124
Table 19. "Rising" category diffusion percentages 125
Table 20. Propagation Patterns 128
Table 21. Mean UoI per Domain 143
Table 22. Mean UoI per Domain 145
Table 23. Mean estimated extended UoI per check 147
Table 24. Mean estimated extended UoI per check 148

List of Figures

Figure 1. The ARPA Network in late 1969 32
Figure 2. The ARPANET in late 1977 33
Figure 3. URL requests in 1995 34
Figure 4. Computer and internet access from 1984 to 2000 35
Figure 5. A simple graph 50
Figure 6. Twitter Social Graph 51
Figure 7. Tree layout of a graph 55
Figure 8. Circular tree layout 55
Figure 9. Fruchterman-Reingold graph layout 56
Figure 10. Real time enhanced random sampling algorithm 69
Figure 11. Mean percentage error of RNS 82
Figure 12. Improvement of mean percentage error over time for RNS 83
Figure 13. Mean percentage error of RNS 85-15 86
Figure 14. Improvement of mean percentage error over time for RNS 85-15 86
Figure 15. Mean percentage error of RNS 80-20 88
Figure 16. Mean percentage error of RNS 75-25 88
Figure 17. Improvement of mean percentage error over time for 80-20 89
Figure 18. Improvement of mean percentage error over time for 75-25 89
Figure 19. Posts per Topic 119
Figure 21. Percent Diffusion of posts 120
Figure 22. “New” category diffusion percentages 123
Figure 23. "Rising” category diffusion percentages 124
Figure 24. UoI variance per check 130
Figure 25. UoI variance per check, log scale 131
Figure 26. UoI variance per check 132
Figure 27. UoI variance per check, log scale 133
Figure 28. UoI variance per check 135
Figure 29. UoI variance per check, log scale 136
Figure 30. UoI variance per check 137
Figure 31. UoI variance per check, log scale 137
Figure 32. UoI variance per check 140
Figure 33. UoI variance per check, log scale 140
Figure 34. Propagation Time and UoI allotment 142
Figure 35. Propagation Time and UoI allotment 144
Figure 36. Extended UoI Allotment, “ImgUr” 149
Figure 37. Extended UoI Allotment, “YouTube” 149
Figure 38. Random Diffusion Visualization 151
Figure 39. Layered-based Propagation Visualization 153

1. INTRODUCTION

My initial thoughts, when we first approached Online Social Networks as a field of study, was to map and model their structure based on Greek users. However, our plans were altered, upon examining the data access policies of various social networks. Policies that were shifted towards more restraining data access environments. A shift in data access policies that brought forth the issue of analysing access limiting environments. Such kind of obstacles provided the opportunity to address research problems in a more direct way. Thus, during my Ph.D. studies, my aim was to address research challenges and discover new topics of interest in social networks. In the course of our research, we identified problems, provided solutions, dealt with novel topics and presented future concepts, in the field of Online Social Networks.

Our research contributes to the fields of data mining and sampling, information propagation and online social structure. Since modern data mining and sampling in Online Social Network environments is impeded by strict access policies, we proposed a new method of sampling limited access environments. Although our solution was built with Online Networks in mind, it can be applied in several sampling scenarios. Additionally, we researched information propagation in multiple Online Social Networks, in order to study the ways, topics and lifespan of information sharing within these networks. Our study was the first publications that dealt with multiple network information monitoring. While, my most recent research proposal is shifted towards the introduction of a new social network that will be focused on anonymity, bias elimination, and equality.

In *Chapter 2* we will present a detailed history of online social networking, from the early concepts of Internet till nowadays, followed by some introductory notes on graph theory, as well as on graph sampling and analysis. Our research on Online Social

Network sampling is presented in detail in *Section 2.3*. In *Chapter 3*, we present our study of information diffusion in multiple Online Social Networks. Prior to *Section 3.3*, where readers can find a detailed analysis of our research, we provide a thorough summary of the most relevant published researches in the field of online information diffusion. In *Chapter 4*, we introduce our proposal for a new type of social network. In *Section 4.1* we present social issues that can be identified in most modern networks. Issues that are the main axis of our proposed experiment, as seen in *Section 4.2*. In Chapter 5, we describe our conclusions, contributions and future research directions.

2. SOCIAL NETWORK SAMPLING

2.1. Online social networks

Social interactions are the basis of society: from a word or a sentence, a nod or some form of physical contact, to modern interactions, like an online post or even a tweet. A plethora of choices to spread ideas, to communicate, to entertain and to be entertained, to experience… The list goes on without end. Hundreds of kilometres can be outreached and with a few clicks people can -almost- instantly interact with each other.

The path that brought us to this fascinating stage of social interactions, could be no less than extraordinary. From the creation of the internet, and the early studies of packet switching in 1960s till nowadays, year after year, social interactions evolved. Up to a point, where these interactions carried to online environments, while the online environments themselves were also developing in parallel. Everything started with a handful of people, communicating and exchanging academic information through predetermined terminals. Leading to today's billions of internet users, communicating with the ease of their handheld devices, choosing from a plethora of online platforms. Simple "text only" emails gave way to live feeds with multimedia content. Somewhere in between the advancement of the Internet and its wide adoption, Online Social Networks (OSNs) were born.

2.1.1. INTRODUCTION

An Online Social Network is an online platform where social interactions occur and social relations are manifested. In such a platform, people can connect and converse with commonly minded individuals, friends, family and acquaintances. Education, entertainment, cultures and news information can be exchanged between any two members or even groups of people within these networks. Every type of connection two

people can have, save for the physical one, can be shared through online social networks. However, not everything was straightforward nor easy, in the first few network creations.

2.1.2. THE HISTORY BEHIND SOCIAL NETWORKING

Social Networks have been in the centre of researchers' attention since the dawn of the 20th century. At that time (1900), one of the very first mentions of the phrase "Social Network" was made in a sociology Ph.D. dissertation [1]. Nonetheless, in 1887, almost 15 years before, social groups had been studied and the concept of social network has been conceived in [2] prior to the introduction of the exact term.

A social network is a structure which represents the dyadic relations of social entities. Each social network is defined by its set of entities and the respective set of dyadic relations between them. This perception of a connected social structure, allowed a set of processes to analyse the network per entity or as a whole. Global and local patterns could be identified, influential entities and network dynamics could be discovered. As a research subject, in the beginning it strictly occupied sociologists, anthropologists and phycologists. It remained such until the last decades of the 20th century. During that period, social networks sparked the interest, amongst others, of computer scientists and mathematicians. In the following section, some of the most influential studies around the creation, adoption and analysis of Online Social Networks will be presented.

In 1986, M.S. Feldman studied information flows in organizations [3]. The author referred to possible alterations of electronic mail usage and its communicational efficiency, in the information exchange of organizations. The significance of the newly formed weak communication ties, as noted in [4], suggested that this new communication method would influence organizational socializing and problem solving.

A year later, one of the first studies on internet applications [5] proposed an explanation for the diffusion of interactive media. The definition of critical mass in [6] played a key role in this study. Critical mass was defined as “a segment of the population that chooses to make big contributions to the collective action, while the majority do little or nothing”. The author highlighted two unique properties of interactive media, universal access created by widespread access, and that reciprocal interdependence is entailed by the use of such media. He later emphasized on the nature of interdependence and its use in new technologies and their diffusion.

Authors in [7] described the interplay between the social environment and the application of communication technologies in organizations. They proposed a simple, recursive model for social information processing with two key extensions, mainly as a way to specify the social mechanisms by which, individuals’ perceptions and behaviours with new media are shaped. They also applied communication network concepts based on these social mechanics. Thereafter, they described the individuals’ use of the media, and the way this usage influenced their position in emergent communication networks. They wrapped up their research with a hypothetical scenario concerning voice mail adoption in corporations. A hypothesis that indirectly pointed to Local Area Networks (LAN) in work environments, as an enhancement of workplace communication.

Using network analytical methods, theories of organizational information processing and social influence resolution, the authors of [8] studied the implementation of an electronic messaging system (EMS) in a small government office. The results provided strong support for the role of critical mass (as seen in [6]) in influencing the adoption of an EMS. Furthermore, pre-usage expectations also played a role in forming enduring evaluations of the outcome of an EMS. The authors also discussed the use and impact of computer mediated organizational media in general.

In [9] the authors addressed the belief of effectiveness in face-to-face communications. In those days, telecommunication scientists tried to create rich and varied interactive systems, in which distance would no longer be a factor in communication. As mentioned in the end of the 19th century [10], *"If, as it is said to be not unlikely in the near future, the principle of sight is applied to the telephone as well as that of sound, earth will be in truth a paradise, and distance will lose its enchantment by being abolished altogether*". The authors concluded that through audio and video channels, interactive tools and mechanisms would encourage usage even when physical interaction is possible. Through these tools a system that would change our perception of presence could be achieved.

Researchers in [11] studied the impact of electronic mail in organizations. The rapid expansion of the communication networks and their advantages were discussed. The authors also reviewed how electronic mail shaped and was shaped by organizational structures and processes. They noted, under certain conditions, the diversity of opinions and better contributions made by its users, mainly due to the uninhibited, nonconformist and conflictual nature of users. They concluded that when people communicated electronically, they could participate in more groups, which became more fluid and periphery got more involved. By periphery, the authors described the outer boundaries of each communication group.

An early precursor of online social networks were computer supported cooperative (CSC) technologies. Although the concept of a CSC existed from 1950, its applications were implemented in 1980 and further consolidated in 1990. The structure of a CSC work environment was pretty similar to that of a modern online social network. Users submitted their questions and help providers came back with potential answers. In [12], the authors addressed "help networks" and the characteristics of relationships within

them. Their findings proposed a variety of areas for further research and potential applications, such as the support of help providers, and the availability of CSCW applications and tools. They concluded that the online transition of CSCW tools would enhance productivity and user satisfaction.

Interorganizational Computer Mediated Communication (ICMC) was the research subject in [13], and was expected to become a mainstream communication infrastructure. To a certain extent, it did manage to evolve in even greater ways. The authors, underlined the role of weak social ties in distributed occupational communities and their usage, towards the study of ICMC growth.

In [14] alternate means of communication used by a certain research group were considered. Those included face-to-face encounters, scheduled meetings, electronic email, fax, telephone and desktop videoconferencing. Their aim was, to learn whether elements in existing communication patterns suggested how future systems could be designed or selected, in order to fit the actual relationships of a group. Their findings showed that a multivariate perspective was necessary to analyse media use. In an interesting note, the main communication media for “major emotional support” were unscheduled meetings and emails.

Researchers in [15] described the distributed Personal Connection (PeCo) Meditator. Such meditators were only available in small groups at the time of the research, but had two main connection system drawbacks. The first was that users had to offer their private information, while the second was that negotiation required between both parts, to lead their cooperative connection. Both problems in such a limited organizational network were resolved with the addition of an exclusive agent that improved the privacy and facilitated the recommendation abilities of the original mediator.

C. Haythornthwaite in a later research [16] described the analysis of social networks and their role in resource exchange by correlating the edge and vertices of a graph with the actors of a social network. The analysis of such networks assessed information opportunities in terms of exposure to and control. The aim was to gain awareness of such an information exchange route and improve the information delivery within that route.

A key aspect of [17] were weak ties theories. The authors investigated such concepts in an organizational environment based on electronic mail exchanges. Their focus was on the relationships between information seekers and information providers. Their results showed that in the absence of personal connections with seekers, providers got problems solved. Thus, the organizational culture received a useful information exchange through weak ties.

During the mid-1990s, people widely adopted various electronic forms of communication, such as newsgroups, bulletin boards, conferences and distribution lists. In addition, various groups were created online to help foster friendship or even romance. But analysts debated whether every collection of people could be considered a community, even with the absence of trust, mutual interest and sustained commitment. These controversies, along with the blurring of the social lives, were studied in [18].

In 1997, Barry Wellman in [19] assayed the relationship between computer and social networks. His consideration was, that as a computer network is a set of machines connected by a set of cables, a social network is a set of people connected by a set of social relationships. The author addressed the analysis of such networks and the information within them. Additionally, he related the networks in: communities and work environments, as well as social and computer networks.

In [20] work and friendship ties associated with different kinds of media usage for different information exchange ways were studied, similarly to [14]. However, their research was focused in a university environment instead of an organizational one. In almost every category of information exchange, the use of electronic mail and unscheduled face-to-face encounters dominated. The authors deduced that the closeness of work and friendship ties were independently associated with more interaction. Specifically, greater frequency of communication, led to exchange of more kinds of information and use of more media.

Another study [21], was focused on CMC Loops project and “Babble”, a prototype application of CMC Loops. The initial conception of Loops was similar to a mailing list (but could be created by any user) and provided freedom in the degree of involvement of each participant. The basic idea was that communicating users could be aware of the participant’s activities with respect to the conversation, so that the crowd would entice others to join too. This was a glimpse of the future, as it pointed to modern online networks.

Online communication systems in 2000, had neglected an important property of natural conversations. The fact that was most conversations were opportunistic and had arisen from the awareness of a shared context. As such, users of those online communities had to join predefined rooms with a strong goal in mind. Contrary to that, the authors of [22] believed that effective communication tools should be embedded in software that supports everyday activities. To that extent, they proposed their own communication system, named I2I.

A study in demographics of groups resistant to online interaction is presented in [23], demographic groups that normally did not constitute communities. However, there was potential for forming online ones. As stated before, common interest, needs, goals

and support were some of the “levers” of online communication. The authors also proposed improvements in online community sociability and usability, which should direct the selection and implementation of technology.

Continuing their work in PeCo Meditators [15], the authors presented a new meditator to easily find capable co-operators with the chain of Personal Connections in a networked organization [24]. PeCo Mediator II was a distributed system to deal with email-based PeCo. Their system was also tested in a classroom, aiding problem solving capabilities of individuals without serious problems.

In [25], almost 200 employees and their work groups were studied. They found that both positive and negative relationships were related to individual and group performance. Specifically, centrality in advice and hindrance networks was positively and negatively affected, correspondingly by individual performance, while group performance was negatively related to hindrance network density.

A model that offered an explanation of social network searchability was presented in [26], in terms of recognizable personal identities, as a form of personalization in content and acquaintances. A sets of characteristics was measured along with a number of social dimensions. The model was applicable in many network search problems, such as peer-to-peer networks, pages on the World Wide Web and, information in distributed databases.

Almost three decades of concepts, studies and applications for social networks in computer connected environments. Organizational needs led the research for interconnected employees and departments, a need for communication which later leapt in households and individuals. But within that time frame, online social networks were born as, a term, concept and form. With small but steady steps, everyday living changed by the advent of these networks.

How could possibly the period of Online Social Networks (OSNs) birth be defined, but for the introduction of the term -and its usage- in early or acclaimed studies?

The term “Online Social Network” was first used in a study focused on computer mediated communication (CMC) [27]. The authors presented measures of social networks analysis such as relations, ties, multiplexity and composition. They also studied the characteristics of such networks; range, centrality and roles. Sampling schemes were presented, along with data acquisition and analysis techniques. Concluding, they asked whether CMC intersected with the structure and functioning of social systems, if the hierarchies of organizations have flattened, were virtual communities rebuilding trust, and had the internet diminished the relevancy of personal attributes? Remarks that were to be answered in the following decades.

Collaboration in distant classrooms was the research object of [28]. More specifically, the authors focused on the exploration of a social network and how one could augment problem solving and cooperation between users. Basically, a study that continued the work in PeCo Mediators, as seen in [15], and further proposed a similar matchmaking agent for heterogeneous classrooms preceding the one used in [24]. In hindsight, this was an early – albeit restricted- concept of a recommendation agent.

In the same educational frame, the authors in [29] acknowledged the technology race in educational institutions and presented two online tools to support technological planning. A web-based instrument for concern measuring and a user centric web site designed for communicating technology stories and networking among teaching personnel.

The ways OSNs could benefit organizations were studied in [30]. It was apparent at the time, that huge companies could identify the driving force of influence within the industry by developing a strategy around OSNs. Although the study seemed a directive

for organizations, it identified OSNs as a key strategic resource that was easily accessible and inexpensive.

The way internet and OSNs affected social capital was studied in [31]. The authors observed, based on data acquired in 1998, that online interaction supplemented face-to-face and telephone communication. Heavy internet usage was associated with increased participation in voluntary organizations and politics, while at the same time commitment to online communities was fairly low. Based on that evidence, the authors noted that internet usage was –increasingly– becoming an everyday routine practice.

The connection between virtual and real social ties was also addressed in [32]. Seven, ethnically diverse, residential Los Angeles areas were studied in regards to their offline and online social ties. It was observed that the level of belonging in real communities was proportional to the prosperity of online community interaction. Ethnic differences were less pronounced than expected and data suggested that online ties were established with people of the same ethnicity.

In [33], the issue of health information available to internet users was studied. At the time, half a billion users had a few thousands of sources to acquire health information. In addition, the potential growth and enrichment of such information sources, as well as the plethora of communication choices offered could drastically improve the effectiveness, efficiency and reach of health education. A small number of early health articles were also summarized, providing education and critique in health-related issues.

The emergence of student centred social environments was studied in [34]. Interactive learning and the increasing adoption of computer mediated techniques was also noted. Researches obtained from social science students over the course of a full

academic year showed the importance of online discussion, regardless whether such discussion media were assessable.

Wellman and Haythornthwaite in [35], two of the earliest adopters of the OSN term, studied the internet and its affects in our everyday lives. They recognized that pre-existing social networks, weaved through ethnic lines and guided online friendship ties. They provided an example, where foreigners that discussed about a topic which was geographically dependant, formed loose online ties based on their common living environment.

Online social networks were increasingly recognized as an important source of influence. Specifically, in the use and adoption of products and services. Thus, a new method of online marketing, namely “viral”, was researched in [36]. It was defined as the process where people (with the respective interest), could market products or services to each other. The huge increase in Hotmail users for two years after its free distribution was presented as a successful case of such marketing.

Another interesting aspect of OSNs was -and still is- activism. In 2003, small steps were made to the direction of social activism, such as petition sites and certain acts of disapproval, as in the case of Lotus MarketPlace [37]. Still many questions arose, what were the rhetorical dynamics of OSN protests, how credible was an electronic petition compared to a written one, could massive web protests make a difference, and ultimately does the speed and reach of OSN communication brought the same features to electronic protests?

Concerning the psychology and the behaviour of some individuals in OSNs, Suler in [38] pointed out that they acted out more frequently and more intensely, than they would in their everyday lives. The factors mainly responsible for this online behaviour were: dissociative anonymity and imagination, invincibility, asynchronicity, solipsistic

introjection and minimization of authority. The loss of inhibition in OSNs was also affected by various personality variables.

In [39], solutions were sought for the information overload caused by the wide adoption of the internet. The semantic web, peer to peer computing and OSNs would further impact online interaction and collaboration. As such, a vision of a semantic desktop was presented, one with a windows style desktop which would enhance the individual collaboration and interaction, while the time needed for information discovery and filtering, would be reduced drastically.

Trust had always been the cornerstone of transactions in our society and trust led to reputation. OSNs could be used, as observed in [40], as reputation mechanisms: initially as automated methods of reputation identifiers based on network position, and secondly as filtering tools for rating users. Both functions would be based on the network structure which might lead to more complex issues, such as reputation brokers and markets. These features, not apparent at the time, have been incorporated in several modern online shops and networks.

By 2005, the scientific interest for OSNs was immense. The structure of such networks was studied in [41] and [42]. The first research presented a theoretical model relating geographical data and social network linkage. The authors noted that with their model, the inversely proportional relationship of new connections is highlighted amongst a number of closer people. While in the latter study, a visualization system was designed, guided by ethnographic researches of those online communities. Targeting the end users, that system proved to be fairly usable and could also facilitate link discovery.

In addition to scientific interest, user participation in social networking sites was dramatically increasing in the years up to 2005. Various OSN services allowed millions of individuals to create online personal profiles and share their personal information, with

vast networks of acquaintances. In [43], which is one of the first essays concerning privacy in OSNs, the authors studied patterns of information revelation in online social networks and their privacy implications. They also evaluated the amount of information disclosed and the usage of privacy settings.

Small communities, colleges and universities were slowly adopting the use of OSNs. The exploration and forming of new social bonds amongst college students was studied in [44]. The results suggested that Facebook was mostly used to learn more about real life contacts rather than initiating new connections. While in [45], the use of OSNs from university students and its relation to social capital formation was the research focus, as in [31]. Political affiliations, common interests and location proved to be some of the social linkage "levers".

Recommendations networks and the enhancement of internet search were also studied. A framework for recommendations that supports the creation of new interpersonal relationships was presented in [46]. In [47], the Internet search through OSNs was studied, aiming to provide personalized search results in interested OSN users, while at the same time preserving their privacy.

From 2007 and beyond, analysis of large OSNs became possible, due to the increased public usage and the –almost– unlimited access to OSN data. Two notable studies that featured unseen depth and data volume were [48] and [49]. OSNs like Orkut, YouTube and Flickr were studied in the first publication, while Myspace, Orkut and Cyworld were studied in the second one. In both of them, the graphs studied contained millions of vertices. The authors' analysis was mainly focused in each network's topological characteristics.

While from 2008 and onwards, a large number of studies were conducted on specific OSNs. User experiences, motives and interaction, within Facebook were studied

in [50], [51], [52] and [53]. Twitter was analysed in [54], [55], [56] and [57], in regards to its identity and functions. Functions that went as far as mood prediction, and global event monitor platforms. Even Myspace that was declining at the time had its own merit of publications, with [58], [59], [60] and [61], which addressed its privacy, social linkage and life impact.

In the recent years, researchers explored several new and revisited some old, aspects of OSNs, such as personalization, recommendations and e-learning capabilities of several networks, [62], [63] and [64]. Privacy is still a relevant topic, as well as data access, [65], [66] and [67]. However, the future of research based on OSNs is still unclear. Research topics could range from sharing, multi-OSN users and content, to new affiliation types and forms of OSNs.

2.1.3. THE INTERNET HISTORY

Before analysing the birth and expansion of OSNs as seen from the perspectives of users and researchers we should comprehend the growth of the internet, from its premature form of the ARPA Network (ARPANET) to its current state.

Although many closed networks existed in the 70s, ARPANET was the first packet switching network. The internet protocol suite, known as the Transmission Control Protocol and Internet Protocol (TCP/IP), was conceived and specified in the 1970s during ARPANET's expansion. ARPANET was the result of a computer network plan prepared by Robert William Taylor, based on packet switching.

Figure 1. *The ARPA Network in late 1969*

As shown in ***Figure 1*** from [68], ARPANET consisted of only four nodes in December 1969. But within eight years the network reached Europe via satellite connection, see ***Figure 2*** from [68], and surpassed 100 nodes. A few years later, the military forced the separation of its nodes, leading to a stripped down ARPANET, which would continue to operate until the late 80s, alongside multiple other networks.

Computer scientists in the mid-1970s were already pondering the unification of all these networks. The specification that, for the first time in history, used the word Internet to describe an internetworking environment, was RFC675[1]. Quoting directly from the transcript, “This document describes the functions to be performed by the internetwork Transmission Control Program [TCP] and its interface to programs or users that require its services.”

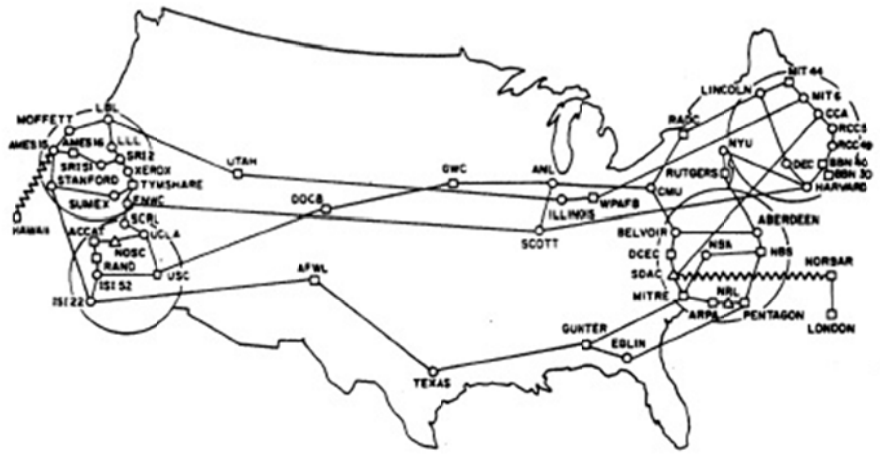

***Figure 2**. The ARPANET in late 1977*

ARPANET interlinked with the National Science Foundation Network (NSFNET) in the late 80s. Consequently, the backbone infrastructure, which started with 56Kbit/s and slowly reached 45 Mbit/s. In the meantime, the European Organization for Nuclear Research CERN was also using RFC675 protocols to connect its internal computers. In 1989, CERN opened its first external connection[2] and ARPANET connected with Australian universities using the aforementioned Internet Protocols.

[1] http://tools.ietf.org/html/rfc675
[2] http://ben.web.cern.ch/ben/TCPHIST.html, Last Retrieved: 13/01/2016

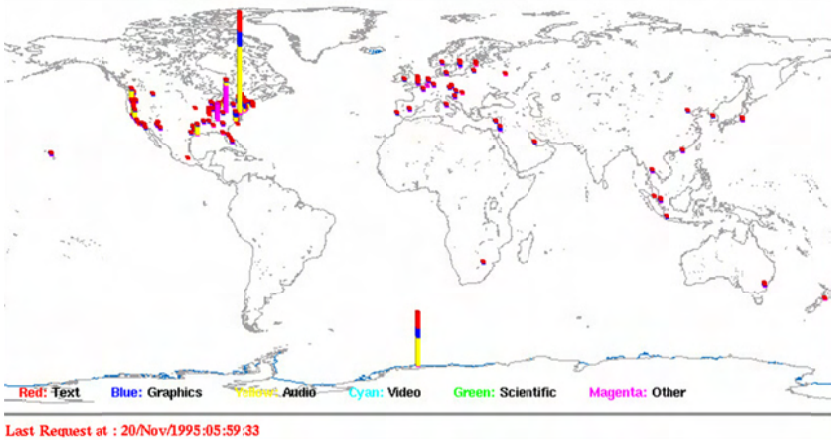

***Figure 3**. URL requests in 1995*

In the dawn of the new decade, every continent had at least one IP-based network. With the help of the traffic visualization application "Palantir"[3], the internet traffic in 1995 was mapped, based on Uniform Resource Locator (URL) requests, see ***Figure 3***. At the time, the only African city with at least some activity was Johannesburg, while all populated continents on earth were interconnected.

Few years later, almost 42% of the United States population had internet access. The corresponding values can be seen in ***Figure 4***, from "Home Computers and Internet Use in the United States: August 2000"[4]. From 2000 to 2010, both personal computers and the internet became more of an everyday need than a specialized tool. Internet users increased exponentially, mostly because by 2010 users' needs for entertainment and socialization could be met by several online services in addition to the existing offline ones. At that period, the frenzy of online social media started. By the end of 2015, more than three billion users have access to the internet, and more than three quarters of its adult users are using social networking sites.

[3] http://archvlsi.ics.forth.gr/html_papers/INET98_Palantir/, Last Retrieved: 13/01/2016
[4] http://www.census.gov/prod/2001pubs/p23-207.pdf, Last Retrieved: 13/01/2016

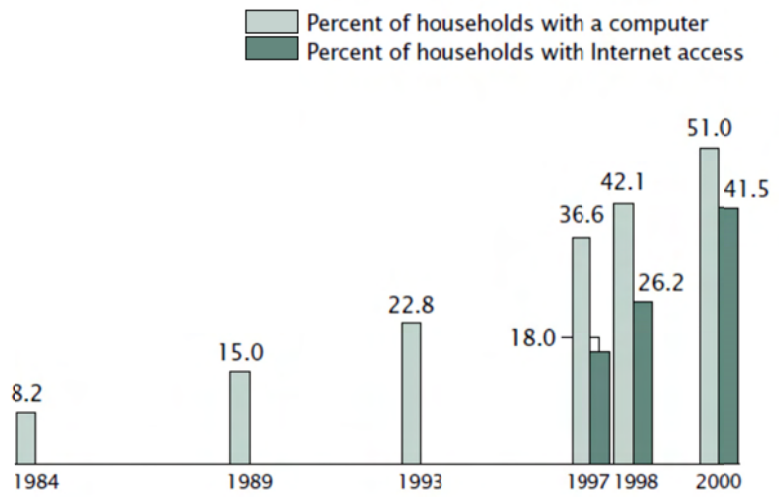

Figure 4*. Computer and internet access from 1984 to 2000*

In the meantime, World Wide Web with its early static nature was slowly replaced by Web 2.0 [69]. Users could now create profiles, interact and contribute to the content and information that was presented to them[5]. Along with the evolution and advancement of World Wide Web, OSNs were born and came of age, taking advantage of the large Internet user base and our existing need for communication. Created more for satisfying needs for professional tools, rather than meeting the demand for a mass media communication platform. Their growth was remarkable and directly comparable only to internet's growth. In the following section we will present a detailed history of how social networks evolved, along with their most successful applications, from a user's perspective. We will focus more on facts and popularity, and not on technicalities and algorithms. The following section is loosely based in the following sources: [70], "The history of social networking"[6], "A brief history of Facebook"[7], "A rundown of Reddit's history and community"[8].

[5] http://www.ibm.com/developerworks/library/ws-socialcollab/, Last Retrieved: 13/01/2016
[6] http://www.digitaltrends.com/features/the-history-of-social-networking/, Last Retrieved: 13/01/2016
[7] http://www.theguardian.com/technology/2007/jul/25/media.newmedia, Last Retrieved: 13/01/2016
[8] http://thenextweb.com/socialmedia/2011/10/14/a-rundown-of-reddits-history-and-community-infographic/, Last Retrieved: 13/01/2016

2.1.3.1. Early Forms of Online Social Networks

Computers in the 1970s and early 1980s were a rare commodity; the need for distant communication was satisfied by telephone and conventional mail services. Yet, it was in 1971 when the first electronic network message was sent by Ray Tomlinson[9]. This message was exchanged between two ARPANET terminals. Although the networked computers were side-by-side, it was a huge step towards communication without range limits.

Community Memory, introduced in 1973, was a computer network of social hubs with more than a hundred terminals placed in Berkeley, California [71]. Through each terminal, users could leave messages that could be read from another connected terminal. It was mainly an experiment aiming to determine public's reaction to computerized communication. However, users utilized the service as a general communications medium for chatting and exchanging news or even for commercial purposes.

The Bulletin Board System (BBS) was the online equivalent of Community Memory. The first BBS started operating in 1978 and besides the obvious difference in online functionality, users could also share files apart from text messages. Unfortunately, because of the telephone rates applied to the connected users, these boards were usually local. Nevertheless, with the advent of the Internet, these local boards were connected through various services. Their peak in both usage and public interest was achieved around 1995.

A similar online service, that allowed users to share news, events and files, was CompuServe. It was one of the main online platforms in the 1980s, broadening the user base of online services. However, its downfall was the high applied charges which were

[9] http://openmap.bbn.com/~tomlinso/ray/firstemailframe.html, Last Retrieved: 13/01/2016

based in hourly rates. CompuServe almost disappeared when America Online (AOL) provided its community services with a fixed monthly subscription, in the mid-1990s.

Another similar online service, highlighting communication rather than sharing, was the Usenet. Established in 1980, its users could read and post messages to various newsgroups. Conversations were in the form of threads and later stored in a server. The main difference from preceding solutions was its decentralization.

However, the older direct ancestor of OSNs, was probably AOL, through which users could create profiles, communities and friend lists, search for friends, browse other members' profiles and communicate with them. AOL was the leading provider of online services for seven years, from 1995 to 2002.

Match.com was created in 1993 and launched two years later. Although it first started as an advertisement system, it went on to become the most successful online dating service. During its first steps, the company tried to persuade initial users with free lifetime rewards. Within a few years, the value of the site was in the millions dollars. During 2004, the network had more than 50 million users worldwide. The number of users has been falling slowly since then but the site has remains operational.

Claasmates.com, created in 1995, was the first service trying to encourage networking with old classmates. However, its subscription-based model and its inability to attract new users were the roots of its progress stagnation. While Classmates.com remained operational, its purpose was slightly shifted towards nostalgia (of movies, music, posters etc.), rather than in interconnecting with old acquaintances.

In 1998, one year after its founding, SixDegrees.com was the first online service to allow its members to search other people's friend lists. Although every other feature seen in SixDegrees.com was already used in some form, in previous online social platforms, their combination and the addition of searching through a friends list made it

the first of its kind. Unfortunately, the reasons of decline were –according to its users– the lack of activities after accepting a friend and –according to the creator– that the public was not ready for such a kind of networking. The service was terminated in early 2000.

In late 1990s, a plethora of sites with OSN features went online and several existing sites added them to their existing structure. Some of them were, BlackPlanet, LunarStorm, AsianAvenue and Livejournal. Amongst them, Cyworld, a famous virtual-world site in Asia, was launched in 1997 and added social network features in 2002. It was one of the first companies to profit from the sale of virtual goods. At that time, several modern OSN tools were added along with several virtual world personalization options. As of 2007, Cyworld had more than 20 million members[10]. Sadly, four years later, hackers stole personal information from the OSN. That hacking event resulted in a feeling of insecurity and distrust from the users. Page views and unique visitors of Cyworld gradually decreased and in early 2014 the majority of users had left joining other social networking services.

2.1.3.2. Modern Online Social Networks

Beyond the concepts and implementations of early online user communication sites, OSNs essentially appeared in the early 2000s. Based on Ryze.com, which was a tool for professionals aiming to expand business networks, Friendster was created. Ryze.com started in 2001 but failed to attract a significant user base. On the other hand, Friendster started as the general public equivalent of Ryze designed to compete with dating sites. It was designed to promote friend of a friend connections with the thought that people in proximity to one's social circle would be better partners. Before the site's

[10] http://research.microsoft.com/en-us/um/people/thomkar/osn/moon_cyworld.pdf, Last Retrieved: 13/01/2016

unveiling in a dedicated press conference, it already had three hundred thousand active members.

After its abrupt recognition and the subsequent impact in users' activity and signups, Friendster encountered technical difficulties. Its rapid growth was not expected and these difficulties resulted in down time and user frustration. Quaintly, because of the exponential user growth, the social context collapsed, friends were mingling with colleagues and former classmates. One of the most prominent differences from today's OSNs was; the inability to check a member's profile if they were a fifth degree or higher connection. This restriction forced users to add seemingly unknown individuals to their network, in order to increase their reach, defeating the purpose of friend-to-friend connection.

An impressive community effect was the creation of fake profiles, either for entertainment or for functional purposes. These fake profiles created for entertainment were either celebrities, companies and similar entities, while on the functionality case, they were universities, campuses or colleges. From the perspective of the hosting company, these fake profiles were harmful to the community and should be deleted. That led to the separation of the company from its users, as they believed that these profiles served their best interests. Most of the early adopters abandoned their profiles and Friendster's popularity started sinking in the U.S.A. In a strange way, at the same time, interest grew abruptly in southern Asia. The ultimate reason of failure was a mix of technical difficulties, social collisions and the company's distrust towards its users.

Around the period of Friendster's user decline, many OSNs launched. Orkut, Flickr, LinkedIn and Myspace, all launched at that period, just to mention the ones that withstood time. Broad audiences, as well as professionals, could choose from a wide range of OSNs. Whilst several sites were created that promoted connections, based on

common interests and activities. Such were, Care2 for activists, Couchsurfing for travellers and MyChurch for members of churches. During that time, an important feature of our time, sharing, was popularized with the likes of Flickr, Last.Fm and a couple years later with Youtube.

2.1.3.2.1. MySpace

Although in the internet world, OSNs were gaining attention, the public and corporations outside the computer industry, were still unaware of such interconnected networks. A great example is MySpace, launched in 2003, which -according to its creator- wanted to compete with Friendster and other similar networks. While they tried to attract previous Friendster users, the online community acted on its own. Most notably, when a rumour that Friendster was going to adopt a subscription model went wild, users began promoting MySpace as an alternative free OSN. This resulted in a rapid growth, mainly by capitalizing on competitor's faults.

Several independent music bands, along with their fans. had also adopted MySpace as their communication medium and as a showcase for their projects. Furthermore, full HTML page customization for user pages was supported, one of the earliest forms of personalization in OSNs. Young internet users, massively started joining MySpace in 2004. At the time, users could be separated into three groups, artists, teenagers and post college adults. Although the interaction between adults and teens was minor, their connecting link was music.

However, when the mainstream media gave attention to MySpace, mainly due to the huge figures surrounding its sale in 2005, several problems occurred. Apart from the security of the network, which raised several safety concerns since it was fairly basic, the reported sexual interactions of adults and teens became a serious matter. MySpace had its peak in 2008 –with roughly 75 million unique visitors per month- but rapidly lost users,

resulting in several takeovers[11]. It was finally sold in 2011 for less than a tenth of its 2005 value.

2.1.3.2.2. LinkedIn

LinkedIn was one of the first business-oriented OSNs started in 2003. It focused on creating personal networks based on occupation, skills and employment. It attracted 4,500 users during its first month of service. The network only allowed users to connect with someone they shared a real-life professional connection or someone they had common acquaintances with. These restrictions where designed to enforce a level of trust to connections. In addition, users that spam invites or send many fake ones, were rebuked or even removed. As of 2015, the network has more than 390 million users in over 200 countries, out of which, over 35 million are students or recent graduates. Moreover, 3 million companies have joined the service, performing billions of professionally focused searches per year[12].

2.1.3.2.3. Facebook

In 2003, the predecessor of Facebook, Facemash was launched as a site only available to the students of Harvard. It used pictures of students and users could vote on who is better looking. It may sound far from what we currently identify as Facebook, but the fact that the pictures of individuals along with their names were used in it could be seen as the inspiration for modern Facebook.

A year later, one of the Facemash creators started working in a new project, theFacebook. Once again only Harvard members were allowed to join in the early days. Within a month almost half of the undergraduate students had joined the network. In

[11] http://www.adweek.com/socialtimes/facebook-crosses-130-million-global-active-users-closes-us-gap-with-myspace/215421, Last Retrieved: 13/01/2016
[12] http://press.linkedin.com/about/, Last Retrieved: 13/01/2016

March 2004, students from Columbia, Yale and Stanford could also join the network, and the expansion continued in America and Canada. A year later, theFacebook omitted “the” from its name, transferred to its current domain and expanded in United Kingdom, Mexico, Puerto Rico, New Zealand and Australia.

The site opened signups for everyone over 13 old years with a valid email address in 2006. It also started hosting company pages, which surpassed 100,000 in 2007. Surprisingly, by the end of 2012 it had more than a billion active users. Whilst it was providing many advertising opportunities from 2007, in 2012 Facebook surpassed one billion dollar profits, thanks to its huge user base and traffic. Despite the lower growth, Facebook is still getting larger and constantly improving. By enriching its available features or by adding new ones, keeping its design fresh and by improving its algorithms. It is almost certain that it will remain the most visited OSN for years to come.

2.1.3.2.4. Microsoft and Yahoo

Despite the fact that MSN spaces was launched as a modern OSN, it failed to keep up with the times or rather with its users’ needs. It started in 2004, with blog support, communication and customization features, such as sharing of photos, commenting, access rights for specific contacts etc. Then, the Windows Live platform launched, a cat that forced a rebrand of MSN spaces to Windows Live spaces. The public interest peaked in 2007, with 27 million unique visitors per month. Unfortunately, Windows Live Spaces closed in 2011, forfeiting to the competition. Microsoft, after the acquisition of Skype, focused more on Skype’s expansion as a communications tool, rather than its enrichment with social network features.

Likewise, Yahoo had a go with online social platforms through Yahoo 360. In it, users could create their sites or blogs, connect with each other, or maintain a profile along with several other OSN centric features. It also had an invitation only period, in 2005, but

later it became available to the public. Its growth halted in 2007, around the same period with Microsoft's service, possibly due to Facebook's rapid expansion. Two years later, Yahoo decided to stop supporting the platform, which was in beta phase, in order to focus development in its successful services. Interestingly, Yahoo Asia launched a similar service in 2008, with the name "Yahoo 360! Plus Vietnam", which remains operational and has reached more than one million users.

2.1.3.2.5. Tencent Qzone

Qzone is the most successful OSN in China. More than 600 million users were active during 2015, at least once per two weeks. Qzone started in 2005 as a blog creation platform, but added many social features which were linked to users' QQ profiles. its successful messaging platform. Through the already existing user base of QQ, Qzone increased its participation exponentially. Apart from the usual OSN features, Qzone offered free and paid customization options and items for the users' profile pages.

2.1.3.2.6. Flickr and Youtube

User-generated content was and remains an important factor and a key for success for the design of a modern OSN from the early 1990s till today. Two major contributors for the integration of user-created multimedia content and its subsequent sharing in modern OSNs were Flickr and Youtube. Both ended in the hands of large corporations (Yahoo and Google respectively) and grew substantially. Despite the fact that similar services existed before them, the success these two enjoyed intrigued OSN developers. Users embraced this type of service and in turn pushed their recreation and integration into several OSNs.

A web-based multiplayer game and its tools were the foundations for Flickr, the popular image sharing site. Even though it could not be put in the same category as

previous OSNs, its character and design choices, heavily influenced every major OSN. Its earliest 2004 version was more of a photo sharing and chatting service, rather than an OSN. Its evolution is obvious in Flickr data limits, from a few MBs of free data upload and storage, to one TB free storage for each user and unlimited data upload. Nowadays, more than 7 billion photographs are uploaded in its servers. After many additions and improvements, Flickr became a modern OSN, where users can form connections, interact, comment and create personal profiles and pages.

Similarly, Youtube, which started in 2005, popularized video sharing throughout the internet. As with almost every modern OSN, in its first days it was more of a "share and view" multimedia platform. However, within a year after its launch and due to its great success several aspects of its design were revisited. It went on to become the third most visited site, behind Google and Facebook, in 2010. Three years later, Google enforced the linkage of the Google plus and Youtube accounts, creating an OSN mixture of a traditional OSN with a video sharing network. We should note that several –not all- OSN key features were already present in Youtube, long before the merge of these networks. In 2014, the number of unique visitors per month reached 1 billion, while more than 400 years of video time are scanned daily[13].

2.1.3.2.6.1. Instagram

The first OSN that was created with mobile users in mind was Instagram. It was made available to the public in 2010 and was focused on photography networking. Users could post their pictures, with or without a caption, and tag them with related words. Of course, OSN features, follow and comment were also there. After submitting their content, users could easily share it to any of the supported OSNs. Instagram had more

[13] https://www.youtube.com/yt/press/statistics.html, Last Retrieved: 13/01/2016

than 200 million users with more than 50 million joining in the last six months, hosting over 20 billion pictures[14], Instagram was bought by Facebook in 2012.

2.1.3.2.7. Reddit

In 2005 Reddit was launched. Reddit was dubbed -by its creators- as "The front page of the internet" and could not be described as a strict OSN. Users could communicate and befriend each other, but they did not have the ability to create a personal profile. The main purpose of this particular social network, was to promote interesting user submitted content. Whether it was news, an image or a simple link, users voted up or down, indicating their approval or disapproval. Content was then moved up (or down) in a serialized and timely order, towards the front page of the site. Users also could -later- create and self-moderate various subcategories, called subreddits. The most popular subreddits were funny, AdviceAnimals and pics.

When it first started, most of the hosted links were fakes and users could not add comments to any of them. Eventually in 2006, Reddit added comment and multilingual support and unique visitors reached half a million per day. In 2008 the ability to create a subreddit was given to the users and the Reddit project became open source. In early 2011 more than one billion pages were viewed per month and before 2012, almost one and a half billion page views were achieved[15].

2.1.3.2.8. Twitter

Twitter is another important but slightly different OSN. Created in 2006, its design was based in short text messages exchanged between small groups. Although its start was not slow, it was neither impressive. Its growth remained stable up until 2007, when within a week its daily usage tripled. Its short messaging texts of up to 140

[14] http://blog.instagram.com/post/80721172292/200m, Last Retrieved: 13/01/2016
[15] http://www.reddit.com/about/, Last Retrieved: 13/01/2016

characters proved a success during a certain conference, where the attending public used the platform to communicate, report news and chat. Blogs reproduced the success of this new OSN in that conference and the public immediately followed that new communication service. The most prominent features, apart the short messaging, was the follower/following feature and the reply function. In a sense, Twitter replicated the act of conveying your thoughts, events, emotions and conversing, with a short multicast message. These effects were even stronger when communicating with celebrities or politicians.

Before the end of 2007, more than 400,000 tweets were sent every day. In February 2010 the 100 million mark of tweets per day was passed. It was apparent that the public had fully adopted this new network. During the same year, a new version of twitter started rolling out, which allowed for the creation of a central page profile. This version reminded of MySpace customization, establishing Twitter as a formal OSN, by offering the essential social features to its users.

2.1.3.2.8.1. Tumblr

Tumblr is a similar to Twitter microblogging social network but less successful. It was launched in 2007 with more than 75,000 users joining within the first two weeks. Aside from text, users can post multimedia content or comment to another post. They can also follow, favourite or reproduce any content posted on the site[16]. The site currently has almost 200 million visitors per month and hosts more than 190 million different blogs.

2.1.3.2.9. Google Plus

Google had its own merit of OSN attempts. From 2008, developers at Google were trying to create a social network to compete with the rival OSNs. Their most

[16] http://www2.technologyreview.com/tr35/profile.aspx?TRID=953, Last Retrieved: 13/01/2016

successful try was their fourth, introduced as Google Circles - later renamed to Google + (Plus). The initial name wanted to highlight the ability to form specific social circles in order to interact and share content with each circle independently. During its beginning, in a private beta phase, only invited users could participate and in turn send a small number of invites. These restrictions sparked the interest of users, who massively joined the network when it went live in September 2011.

Within three months, over 50 million users had joined the service, making Google Plus the fastest OSN to reach 50 million users. Even though more than a billion users have signed up, the current number of active users is around 300 million, which establishes Google plus as the third largest OSNs by active user numbers, only behind Facebook and Qzone. Considering it is a relatively young network, no one knows how it will develop. Its growth has currently slowed down, as it happens with any modern and popular OSN, but it is unknown if it simply has come of age or the public interest will reignite.

2.1.3.3. The present state of Online Social Networking

The discussion for the current state of social networking could be viewed from two seemingly separate viewing angle, namely the user and the company perspective. Users want more personalization options, content diversity, accessibility, security, mobility and the feeling of relevancy to the service, while companies in order to provide and improve those features need to sustain a healthy business model. As we will present, there is not such a long distance between these two interacting groups.

Currently, the main trend in social media design is sharing, since the users want to share, a picture, a video, a song, a destination, a dream, a job position or a thought, they want to discuss and publicize it. Sharing occurs some times to a small circle of trust, some other times to everyone interested, and rarely to everyone. Social recognition and

stature is intertwined with OSNs. More tools and platforms that simplify sharing are created daily and every modern site offers these OSN sharing features, at least in some form. Most importantly, with such a sharing trend, a website “outsources” the tasks of diffusion to its users. So, if sharing of every kind of content is favourable to the hosting website, should we be concerned of whether we were forced to employ sharing or we adopted it ourselves? In such an interconnected world, there is a thin line between spontaneous and enforced adoption.

Therefore, a company which would wish for viewership is also positively influenced by the current social media trend. As with most websites, the first and best source of OSNs’ income is advertising. Most of OSNs, after they acquire a significant user base turn to several advertising solutions and tools. However, during the recent years, advertising practices in OSNs are becoming extremely aggressive, sometimes with the tolerance of the social network. In an effort to attract, or to be more precise, lure users to their sites, advertisements are becoming gradually less distinct from the actual content of the site. Most of these controversial cases have occurred in news media sites, but OSNs had their fair share of similar advertisements.

For example, in Reddit, advertisements are in the form of a normal post only recognizable by a light blue background, whereas in Twitter, promoted tweets from businesses are only identifiable by a small “promoted by” in the lower part of each tweet. It seems that without subscriptions, ad revenues are the only reliable source of income for modern OSNs. Should this affect user experience, or to what extent users will remain faithful to their preferred OSN, after the inevitable penetration of “camouflaged ads”, remains to be seen.

Finally, two very important but often overlooked issues of modern OSNs, are handling of personal data and privacy. How secure is user data and can it truly be

private? To answer this question, we have to look at some disheartening facts. In 2012, 6 million passwords from LinkedIn leaked online[17], while in 2013 more than 250,000 Twitter user profiles were hacked[18] and at least 1.5 million Facebook users had their personal information put up for sale[19]. Apparently, not even the biggest OSNs can reassure users that their information is safe from malicious attacks. The irony is that their privacy is not only endangered by external threats. As was expected, since it has the biggest user base, Facebook had raised the most reported privacy concerns. From the inability to completely delete your profile automatically[20], where information is stored in case you want to come back, to the granting of personal information to third party advertising companies without any authorization from the user[21]. On the other hand, Twitter shamelessly sold a collection of more than 300 million tweets to advertising and data solutions companies[22], under the pretext that they were going to be removed from their database. Thus, it seems that users should be aware of the risk they take when publicizing personal information in OSNs. Once online, no one can guarantee the absolute security and privacy of the uploaded data.

[17] http://money.cnn.com/2012/06/06/technology/linkedin-password-hack/, Last Retrieved: 13/01/2016
[18] https://blog.twitter.com/2013/keeping-our-users-secure, Last Retrieved: 13/01/2016
[19] http://www.computerworld.com/article/2517490/security0/1-5m-stolen-facebook-ids-up-for-sale.html, Last Retrieved: 13/01/2016
[20] https://www.facebook.com/help/224562897555674, Last Retrieved: 13/01/2016
[21] http://www.forbes.com/sites/anthonykosner/2013/08/31/new-facebook-policies-sell-your-face-and-whatever-it-infers/, Last Retrieved: 13/01/2016
[22] http://rt.com/news/twitter-sells-tweet-archive-529/, Last Retrieved: 13/01/2016

2.2. Online Social networks as graphs

As mentioned above, OSNs provided a great source for both social and mathematical graph analysis. However, due to their constant expansion and their dynamic nature, as demonstrated in *Section 2.3.2*, advanced methods and lots of resources are required for processing their statistical properties, which are constantly becoming more complex in every research subject, from simple visualization and sampling to semantic analysis of graph content. Before analysing existing research in graphs of OSNs, let us first provide the basics around graphs.

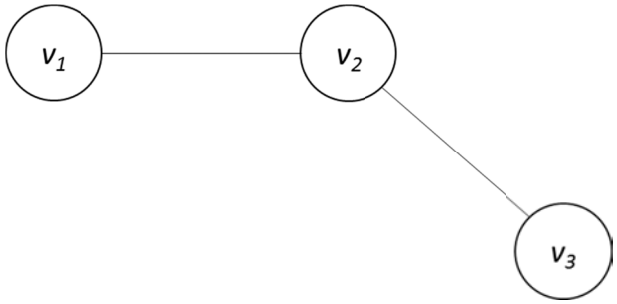

Figure 5*. A simple graph*

2.2.1. GRAPH THEORY

A graph is the representation of a relation between two objects. The basic visual depiction of an undirected graph can be seen in ***Figure 5***, and a more advanced visualization of a multi component undirected social graph can be found in ***Figure 6***. The visual perception of a graph, although fairly simple, is based on the mathematical definition of its structure and properties. The mathematical study of graphs, namely graph theory, was introduced by Leonhard Euler in 1736. His paper [72] around the Seven Bridges of Konigsberg problem marked the beginning of graph theory.

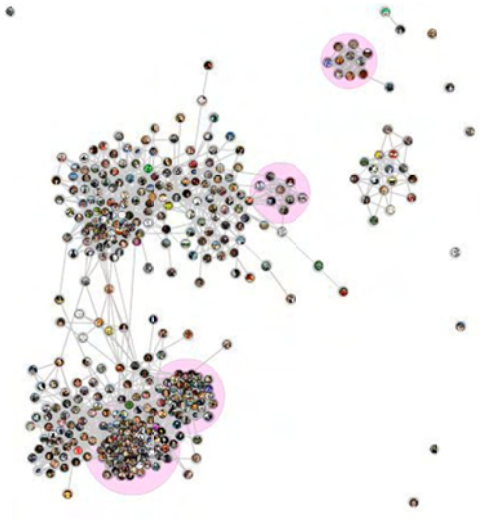

***Figure 6**. Twitter Social Graph*[23]

In mathematics, a graph is a structure which models the relation between objects. These objects are called nodes or vertices (node and vertex in singular). The relationship between any two vertices is called an edge or a link. Thus, the mathematical formulation is as follows. Let *V* be a finite set of vertices, then:

$$E(V) = \{\{u, v\} \mid u, v \in V, u \neq v\}$$

is the set of relations of two elements of *V*, where *u* and *v* represent these two elements. A graph *G* is defined as a pair:

$$G = (E, V)$$

The elements of V are the vertices of graph *G* (written as V_G) and the elements of E are the edges of *G* (written as E_G). A pair of vertices $\{u, v\}$ is usually noted as uv. When $uv = vu$, the graph is undirected. The number of vertices $|V_G|$ is called the order of *G* and the number of edges $|E_G|$ is called the size of *G*. For an edge, $e = uv \in G$, *u* and *v*

[23] http://www.technollama.co.uk/is-twitter-bringing-us-closer-toghether, Last Retrieved: 13/01/2016

are called its ends. Two vertices *u* and *v* are adjacent or neighbours, if $uv \in G$. As such, in ***Figure 5*** we can see the visual representation of a graph *G* with $V_G = \{v_1, v_2, v_3\}$ and $E_G = \{v_1v_2, v_2v_3\}$. Obviously in this graph $|V_G| = 3$ and $|E_G| = 2$. Vertex v_2 is a neighbour of both v_1 and v_3, but vertices v_1 and v_3 are not neighbouring with each other.

When a graph is allowed to have loops vv and multiple edges between two elements of V, then that graph is called a multigraph $G = (E, V, \psi)$ where $E = \{e_1, e_2, \cdots, e_m\}$ is a set and:

$$\psi: E \to E(V) = \big\{\{u, v\} \mid u, v \in V\big\} \cup \{vv \mid v \in V\}$$

is a function that assigns a pair of vertices to an element of *E*, $e \in E: \psi(e) = uv$. While if in any graph (simple or multigraph) $uv \neq vu$, then the graph is defined as directed. A graph *H* is a subgraph of a graph *G*, if every vertex of *H* is in *G* ($V_H \subseteq V_G$).

The degree of a vertex *v*, $deg(v)\ or\ deg\ v$, is the number of neighbouring edges. For our example in ***Figure 5*** $deg(v_2) = 2$, $deg(v_1) = 1$ and $deg(v_3) = 1$. The degree sum for any graph *G* is:

$$\sum_{v \in V} deg(v) = 2|E_G|$$

and the mean degree of a graph is:

$$\overline{deg(v)} = \frac{2|E_G|}{|V_G|}$$

This formula gives in our graph example, $\overline{deg(v)} = 1.333\ldots$, since we have $|E_G| = 2$ and $|V_G| = 3$. Apart from the mean degree of a graph G, we also have the maximum degree of graph *G*, $\Delta(G)$, and the minimum degree of the graph *G*, $\delta(G)$.

The global clustering coefficient of a graph is a measurement of how closely clustered are the vertices of the graph. It is calculated by the average of local clustering coefficient of each vertex. The local clustering coefficient of a graph is given by:

$$C_{v_i} = \frac{number\ of\ triangles\ connecting\ with\ vertex\ i}{numbers\ of\ triplets\ centred\ on\ vertex\ i}$$

where a triplet is a set of two vertices connected to vertex *i* and a triangle is a set of three vertices connected in a way that they form a triangle. When the number of triplets centred on a vertex *i* is equal to zero, which means vertex *i* is an isolated vertex, then its local clustering coefficient is also zero and the above formula cannot be applied. As said, the global clustering coefficient is calculated by the average local clustering coefficient of every vertex:

$$C = \frac{1}{|V_G|}\sum C_{v_i}\ , \forall v_i \in V_G$$

and the maximum value of both the local and the global clustering coefficient is 1. Note that when a graph has a global clustering coefficient of 1, we have a complete graph, where every vertex (v_i) is connected with a unique edge to every other vertex $(v_j, for\ j \neq i)$.

On the other hand, the assortativity property shows us the depth of similar degree vertices correlation or the different degrees' vertices correlation. Let fe_{ij}, be the fraction of edges in a network that connect a vertex of type *i* to one of type *j*. Then,

$$\sum_{ij} fe_{ij} = 1$$

$$\sum_{j} fe_{ij} = a_i$$

$$\sum_i fe_{ij} = b_j$$

where a_i and b_j is the fraction of each type of end of an edge, which is attached to vertices of type i. In an undirected graph, it is apparent that $a_i = b_j$. Assortativity is then defined as:

$$r = \frac{\sum_i fe_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

where $\sum_i fe_{ii}$ is the sum of the fraction of edges (fe_{ii}) that connect two vertices of type *I*, while $\sum_i a_i b_i$ is the sum of the fractions of each type of end of an edge (a_i and b_i), which is attached to vertices of type i, with $-1 \leq r \leq 1$. When $r = 0$ the graph is non assortative, when $r = -1$ the graph is perfectly dissortative and when $r = 1$ the graph is perfectly assortative [73].

Concluding the section of basic graph terms and definitions, we should address graph components property, which is defined as the number of components a graph is comprised of. Noteworthy, a single isolated vertex of a graph is considered a separate component. Furthermore, there is a lot more depth in graph theory, concerning undirected, directed, weighted and even more types of graphs and properties that cannot be briefly covered. The definitions provided in this section cover every attribute presented and used throughout *Chapter 2*.

2.2.1.1. Graph Drawing

A common interest area for mathematicians and computer scientists alike, is graph drawing. Borrowing elements from information visualization and graph theory, the goal of graph drawing is to clearly visualize a graph structure with respect to the graph's properties and desired applications.

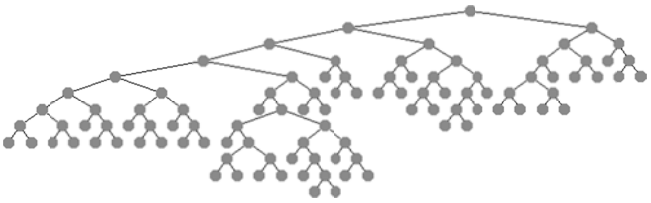

Figure 7. *Tree layout of a graph*[24]

The most common drawing method for a graph is the tree layout. In tree layout a vertex is usually drawn on the top and each neighbouring level drawn towards the bottom, as seen in ***Figure 7***. This method, besides being easy to draw and understand cannot be used for large graphs since its layout will not highlight potential key properties. Yet, many different variations have been proposed, with the best known being the circular tree, as seen in ***Figure 8***, *w*here instead of drawing each neighbouring level of vertices below the initial vertex, each level is drawn away from the initial vertex and is inscribed to a circle.

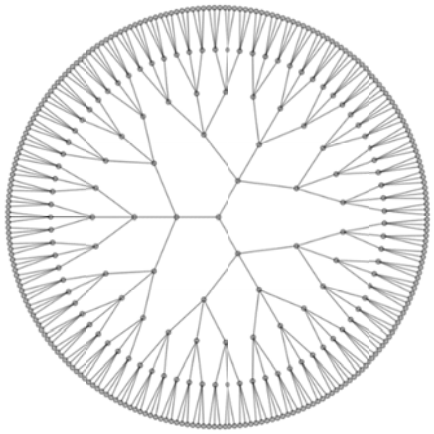

Figure 8. *Circular tree layout*

[24] http://www.informatik.uni-koeln.de/ls_juenger/uploads/RTEmagicC_11562d3e53.gif.gif, Last Retrieved: 13/01/2016

In OSN graph drawing, we need to highlight any important properties, such as the clustering coefficient and the vertex degree. The methods and the required procedures that allow such kind of drawing are called Force Directed Graph Drawing algorithms. The most established algorithm of this type is the Fruchterman-Reingold algorithm, which was presented in 1991 in [74] and quoting its authors: “Our heuristic strives for uniform edge lengths, and we develop it in analogy to forces in natural systems, for a simple, elegant, conceptually-intuitive, and efficient algorithm”. Its step-by-step process replicates forces, like pull and push, which improve the layout of the graph until it reaches a user-defined threshold. As seen in ***Figure 9***, the goal of distinction is achievable with this algorithm. Elements like clusters, power vertices and isolated vertices are pretty easy to identify.

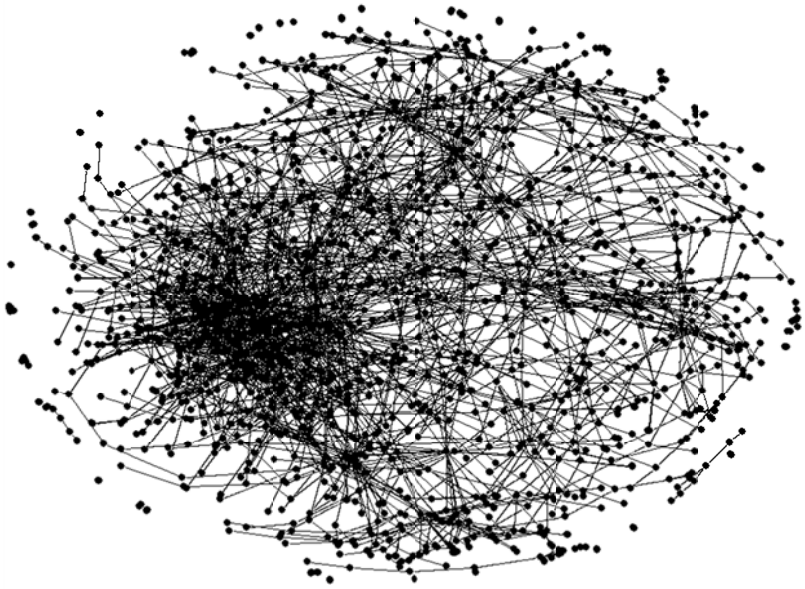

Figure 9. *Fruchterman-Reingold graph layout*[25]

These are the most common, if not basic, graph drawing techniques. Many different options exist, like cluster colouring, vertex size drawing related to degree and so

[25] http://www.biomedcentral.com/content/figures/1471-2105-10-19-5-l.jpg, Last Retrieved: 13/01/2016

on. Moreover, directed graphs, also present a different challenge, since in and out edges must be taken into consideration. We encourage interested readers to take a look at a popular graph visualization platform, named Gephi[26], where they can test the Fruchterman-Reingold algorithm along with the most recent Force Directed Graph Drawing algorithms.

2.2.2. GRAPH SAMPLING

Graph sampling is the technique by which we obtain a set of vertices and edges, essentially a sample, from a graph. The obtained set can usually form a subgraph of the original graph. The graph we took the set from, is mentioned as the original graph, while the subgraph that is formed by the sampled vertices and edges is called the sample graph. Many different sampling methods exist. Most of them work on specific cases and no sampling algorithm provides universally good results. The sampling results are evaluated based on the provided values and properties of the sample graph, compared to the original graph. The methods and their evaluation presented are focused on OSN graph sampling, despite that some of them were not initially created for OSN graph sampling.

The most common graph sampling method is Random (Node) Sampling. In this method, the starting vertex is –usually– randomly picked. The algorithm then randomly picks any other vertex. And so on, until we reach a required sample size. Although it is a widely used sampling method, it fails to accurately project most graph properties [75].

Many variations and improvements proposed, try to enhance social graph sampling, using the core random sampling method [76]. Most notably, the Random Edge Sampling and the Random Node Edge Sampling are both variations of the initial Random Sampling. In the first method, an edge is randomly sampled along with its both ends, then

[26] http://gephi.github.io/, Last Retrieved: 13/01/2016

another random edge is picked. In the second method, a vertex is the starting sample object, followed by a randomly chosen adjacent edge and its end vertex. The authors of [76] note that the first method is biased towards high degree vertices, since most edges are connected to them.

Some other, less popular, sampling methods are the Random Node Neighbour, the Random Walk and the Random Jump. In Random Neighbour Sampling the first vertex is randomly picked along with every neighbouring vertex. In the next step another vertex is randomly picked, not necessarily a neighbour of the previous vertex. In Random Walk Sampling, after the initial vertex is picked, a walk is performed on the graph, adding every vertex and edge traversed. In every step, a probability exists that will reset the walk to the initial vertex. A variation of Random Walk Sampling is Random Jump, where everything works the same, except that the reset probability picks a random vertex instead of the initial one. This modification prevents the algorithm from getting stuck or confined in a single path.

Forest Fire Sampling is another algorithm focused in huge graphs. It was presented in [77] and was basically a combination and improvement of the Random Node Edge Sampling and the Random Walk Sampling algorithms. The first vertex was randomly picked and then the “burning” of adjacent vertices started. With a specific parameter, namely the forward burning probability, an adjacent edge was burnt along with its endpoint. That endpoint was the next vertex the fire spread to. A backwards burning parameter also exists in order to replicate the rekindling to a previous vertex.

Breadth First Search (also known as Snowball Sampling) is a sampling method that aims to completely crawl a graph. Not a feasible task in modern OSNs, mainly due to their scale. The algorithm selects a seed vertex and crawls every neighbouring vertex. After crawling the first level of neighbours, it moves to the second level and so on. A

variation of Breadth First, is Depth First Search. Instead of crawling every neighbouring vertex, Depth First Search functions as Random Walk Sampling, until it reaches an end. Upon reaching its “dead end” it backtracks a level and searches for another path. Both of these algorithms are almost useless in billion vertices graphs, mostly because they usually provide biased samples, while incomplete Breadth First Search samples, fail at providing a reliable value for the level of symmetry and power law coefficient [78].

Nevertheless, not every sampling method can be used in Online Social Network sampling. A sampling procedure is in fact a form of web crawling when focused in OSNs. Most of the methods described above require at least some prior knowledge of the network. Such knowledge is impossible to obtain before the sampling begins, because in an online environment graphs are dynamic, constantly expanding and changing. This dynamic nature of OSNs will always be a great obstacle to sampling. Even more in recent years, where data access is fairly limited within such networks. Such restrictive networks were the focus of our OSN graph sampling analysis. But before we define the problem and present our proposed analysis, we should mention the most emphatic work in graph analysis to date.

2.2.3. GRAPH ANALYSIS

Graph analysis has always been an interdisciplinary subject. Although the focus transcended from offline to online networks and graphs, the analysis was (and still is) a significant research topic with many applications, and various approaches ranging from simple metrics to advanced trend prediction. The amount, though, of disseminated information is vast, thus suggesting graph sampling as an important process for both achieving OSNs related goals, as well as analysing an OSNs’ structure and properties.

The first and most persistent finding of graph analysis is power law distribution of degrees, i.e. the fact that a small portion of vertices is the end of a large portion of the

edges. The power law theory was first proposed in the 19th century in [79] to describe the inequality of wealth in Italy of 1896. Since then this distribution has been reported in several studies, like [80], [81], [82] and [83].

2.2.3.1. Related Work

One of the very first deep graph analysis studies was authored by Broder et al. [84]. This research was conducted with data acquired by crawling Altavista. Their study set up the fundamental macroscopic elements and attributes of a large evolving graph, the World Wide Web (WWW). The authors verified strong theoretical concepts and observations, such as the power law distribution, which appeared in macroscopic (OSN graph) and in microscopic (OSN single user neighbouring cluster) scales.

An early attempt to describe some solvable models with respect to the OSN structure based on random graphs with arbitrary degree distributions, appeared in [85], where models for different graphs were provided, and were later applied to real OSNs. It was observed that in some OSNs of those years, the existing models provided sampling instances capable of describing the entire graph, whereas in the majority of OSNs the same methods were not applicable, perhaps indicating a different social structure in the network, which could not be captured by the selected random graph sampling methods.

Later on, in [76] an extensive analysis and evaluation with respect to large graph sampling was provided. Up to those years, the literature related to sampling techniques from undirected graphs found that some graph properties could be preserved through random vertex selection on samples smaller than the 30% of the original graph size [86]. In [76], simple uniform random vertex selection methods was claimed to outperform edge selection-based sampling strategies in both static and highly evolving graph patterns, even with sample sizes less than 15% of the initial graph.

In [49], the authors analysed the structure of a huge graph, its properties, degree distributions and users' behaviour. This analysis was based on the belief that OSN attributes resembled those of real human-based social network. They compared the structures of three online social networking services, namely Cyworld, MySpace, and Orkut, each with more than 10 million users at the time of their research. Having analysed the complete Cyworld network and parts of MySpace and Orkut obtained with Snowball Sampling, they ended up in the conclusion that different types of OSN users were strongly related to large graph attributes, such as to clustering coefficient distribution, assortativity/disassortativity, network size, average path length, and effective diameter. Additionally, the authors evaluated the snowball sampling method as a breadth first search method on a graph with 12 million vertices and 190 million edges.

In [87], the community identification of huge information networks is studied. Many different databases were analysed, with the graph conductance - as described in [88] - playing a key role in conclusions. An important outcome of the work was the fact that, community size was proportional to "blending" with the whole network. The authors also studied the implications and metaphors in the case of graph partitioning algorithms and community detection, in real-world networks.

In the work described in [89], Zou and Holder -by accepting that frequent pattern mining plays a significant role in sampling large graphs- validated the concept of subgraph mining. Through their technique, called "Random Areas Selection Sampling", they handled sampled graphs along with the initial graph, and then compared the obtained results. They used efficient sampling methods for estimating subgraph concentration and for detecting network motifs, as well as sampling approaches in order to reveal reliable subgraphs from large probabilistic graphs as described in [90] and [91]. Zou and Holder in [89] also claimed that their proposal had the highest accuracy among all other graph-

sampling methods. Their study was based in experimentations in large graphs, such as these of Citations, Amazon and WWW conference series. Another important work in subgraph mining was [76], where the authors evaluated several sampling methods (such as random node, random edge, random jump, etc.). Unfortunately, in relation to OSN sampling, these methods required some or full knowledge of the original graph.

Random walk sampling has been an established sampling process in large OSN graphs. The works in [92], [93], and [94] employed random walk sampling methods in order to sample user entities in large OSNs such as Friendster, Twitter and Facebook. The authors in [95] proposed a sampling method capable of exploring multiple dependent random walks to further improve the sampling procedure in loosely connected subgraphs. Similarly to that, the authors in [96] introduced the concept of multigraph sampling by considering the graph as a union of all the single graphs projected by its multiple relations. For this kind of sampling processes, they proposed a novel two-stage algorithm that walked on the union multigraph by first selecting the relation on which to walk, and then by taking into account only the neighbouring vertices, with respect to that specific relation.

Trying to manipulate big data from large evolving graphs or OSNs and to correlate them to fundamental web metrics, Noordhuis et al. provided an insight of mining Twitter data and of the application of PageRank classification in it [97]. The amount of analysed data was vast, thus the authors introduced the use of a cloud computing solution.

Concerning the data collection and crawling procedure, the authors in [98] presented in a short way, the framework of parallel crawlers for OSNs by employing a centralized queue. Their crawlers used Breadth First approach for fast crawling of eBay profiles. At the same year, one of the largest OSN crawling studies was conducted in

[48]. Their measurements confirmed the most well-established large graph properties in OSN environments, i.e. that vertices with high in-degree tended to have high out-degree values, the clustering coefficient was inversely proportional to vertex degree, and that OSNs graphs were held together by about 10% of the vertices with the highest degree.

Based on these works, the authors in [92] focused on Twitter and showed that also within that OSN, most of the well-known properties were also confirmed. Furthermore, by performing experimentations on three acquired sampled datasets, they managed to classify specific OSN entities (Twitter users), evaluating their outcomes with respect to a crawling limiting environment such as the Twitter API.

In [99], authors analysed the crawling process of a graph, mainly the part of edge and vertex discovery, and presented an excellent related work summary in the field of social network crawling and social graph sampling. They also addressed the issues of multiple seed choice and protected users, which affected the crawling and sampling process on an OSN.

Closing our related work section, we have to note also that several studies on graph analysis focused on other disciplines as well, such as graph compression/transformation, techniques used to enhance the computational speed of applied algorithms [100], efficient storage and retrieval of the analysed web-graph [101], and various methods used in size and evolution estimation of large graphs [102], [103].

Most of these studies sparked our research interest for OSN graph analysis. During our initial approach to OSN sampling, new obstacles were identified and our focus shifted from analysing an OSN graph sample, to efficiently obtaining one. In the following section, a possible solution for a particular problem of OSN sampling/crawling and its subsequent analysis will be presented.

2.3. Real Time Enhanced Random Sampling

A deciding factor in our selection and approach, of a suitable OSN and its subsequent analysis as graph, was the type of data we were going to handle. First and foremost, it needed to be up to date. The OSN we would crawl for that data would need to be fresh, present a high level of activity, be commonly accepted as a modern OSN and allow unobstructed access to its data. Twitter was such an environment, considered as a well-established OSN with a strong social impact [55] [56]. Their data access policies were fairly open, and our aim of Snowball sampling a part of it seemed feasible. Our ultimate goal was to sample and analyse the entire Greek Twitter user base graph.

Twitter, as mentioned in *Section 1.3.2.5*, is a popular microblogging service. A user that joins the network can post her/his own messages, with maximum 140 characters per message. In addition, a user can follow other users, or be followed by others. In essence, these follow and following relations form a directed social graph of connections among users. In such a graph every user functions as a vertex of the OSN graph and every connection amongst users as an edge. The followers of a user represent the ingoing edges of a vertex, while the followed users the outgoing edges of the vertex. This is the basic underlying graph structure of this network, which, by all means, is not considering every aspect, but provides a sound basis for graph analysis.

2.3.1. Problems of OSN analysis and sampling

Twitter graph information is available through its Application Program Interface (API)[27]. In the early days or Twitter, its API allowed a virtually unlimited crawling of its edge data. The so called “requests” on its API could reveal a predetermined amount of information. For example, one request could reveal a list of a user’s messages, a list of a

[27] https://dev.twitter.com/rest/tools/console, Last Retrieved: 13/01/2016

user’s followers or a list of the users following him/her. Although for everyone interested the request limit per hour was capped at 350, developers could fill a whitelist form which upon confirmation would increase available request to 20000 per hour. Unfortunately, the whitelisting procedure was discontinued, right before we started deploying our crawlers. This forced us to reconsider our crawling procedure, as well our ultimate research objectives.

For the interest of precision, we should present in detail, the amount of information available per request. One Twitter API request could be used to, obtain the number of followers and following of a user, essentially discovering the in/out degree count of a vertex. Furthermore, through one request we could also obtain the links (followers or following) of a user in sets of 5000, effectively discovering the in/out neighbours of a vertex, 5000 at a time. This practically meant that in order to get the required graph data of a user with 10000 followers and 5000 following, we would use three requests: two requests to discover its followers and one to discover its following links.

The imposition of the 350 request limit and the abolition of the whitelisting procedure, affected our crawling plan. We decided to split the process to multiple user accounts. This way we could get 350 requests per hour, multiplied by the number of accounts we had attached the crawlers to. In addition, the crawling process needed some level of Twitter authentication. The subject of crawling a limiting data access network and the possible ways we can overcome these limits became another important issue.

2.3.2. Proposed Enhancements

The aforementioned Twitter API limitations forced us to reconsider our means of obtaining a graph sample from Twitter in particular, or potentially another OSN that sets limits to its data access as well. Our first thought was, since API requests are a finite

resource, how we can fully utilize them in order to best sample a graph? Followed by how effective are the sampling methods proposed in the last decade, when applied to OSN sampling scenarios? If they are not as effective as in other areas, how can we enhance these methods to further improve our results?

By carefully approaching the aforementioned concerns, we decided to try and improve sampling processes and also test our proposed improvements. The method we chose was the most common sampling method, Random Sampling. Our tests were conducted in 14 sampled graphs we crawled from Twitter, which we will refer to as Test Graphs (TGs), along with some existing graphs from the Stanford Large Network Dataset Collection[28]. Our proposal utilises every bit of information available in a single vertex of an OSN, taking into consideration the limiting request environment and the demand for efficiency.

Our proposed algorithm could be split into two separate parts that could also work independently from each other. During the first part, we analyse vertices but only sample those that are of significance to us, while in the second stage, we enhance the sampled graph structure with the available sampled information. Both these enhancements are executed in real time and in parallel to the crawling process. Although our tests and comparisons are based on Random Sampling methods, the proposed enhancements can be incorporated, with little effort, in every other OSN sampling method. Based on its real time process and its enhancement to simple Random Sampling, our proposal was named, Real Time Enhanced Random Sampling.

2.3.2.1.1. Selective Sampling

Our first enhancement is based on the fact that social networks' degree distributions follow the Pareto Principle [79], which means, as mentioned in *Section*

[28] http://snap.stanford.edu/data/index.html, Last Retrieved: 13/01/2016

2.2.3, that a really small number of users are the one end of the strong majority of connections. As such, we propose to sample a predetermined number of vertices that follow a predefined distribution. Thus, we restrict our sample to a specific degree distribution, while at the same time we preserve API requests. This request conservation is possible due to the selection and rejection of the graph's vertices, instead of the usual Random sampling process where every crawled vertex is added to the sample.

Even though power law distribution of a graph can be estimated by various methods, such as by the Maximum Likelihood Estimation [104] and the Kolmogorov–Smirnov estimation [105], every method requires full -or at least some- knowledge of the graph and its properties. Unfortunately, when crawling Twitter (or any other OSN) in real-time the only known property is the in/out degree of each crawled vertex. By utilising this information, the most effective sampling procedure must be determined.

In the sampled graphs, we decided to force a desired degree distribution in a small set of crawled vertices. In particular, we searched for the top 20^{th} percent of the vertices -based on degree value- accounted with the 80% of the graph's cumulative degree (80-20 distribution). In order to determine which was the best possible OSN sampling distribution we also tested the 85-15 and the 75–25 distributions. This means that in the 85-15 distribution, set of vertices were accepted into the sample only if the top-15% of highest degree vertices had the 85% of the total vertex degree of the set. While similarly in the case of the 75-25 distribution, only sets of 25% of highest degree vertices had the respected 75% of the total degree of the set, were accepted. Essentially, we are forcing our sampling method to obtain a sample with a predefined power law distribution.

2.3.2.1.2. Enhanced Subgraph

The second improvement is based on the observation that when we sample a graph, we only use the resulting subgraph for analysis. By doing so we probably lose

vital and hard to obtain information, considering the limitations imposed. Thus, we propose to add all the neighbouring vertices of sampled vertices to the subgraph. In fact, this is the combination of two techniques: random and neighbourhood sampling. In addition, this will not be resource taxing, since by crawling a vertex, we also discover its degree and neighbouring vertices.

Sampling a graph G provides a number of vertices and a number of edges. In almost every case after completing the sampling process, the next step is to find the subgraph consisting only of sampled vertices with any sampled edges amongst them, followed by analysing the resulting graph. A possible flaw of this process when used in OSNs sampling scenarios is that it discards many discovered edges along with their vertices. These edges and vertices are the discovered first level neighbours of the outboard vertices. We aim to test whether the usage of these vertices and their respective edges is beneficial to our sampling goals.

2.3.3. THE ALGORITHM

In every OSN sampling scenario, sampling and crawling are considered as one process. In order to take a sample from an OSN, a crawling procedure must be performed. By crawling an OSN, vertices and edges are discovered, similar to any graph sampling process. Our proposed sampling method, along with its enhancements, can be separated in two parts. This distinction is apparent in the flow chart of the algorithm shown in ***Figure 10***.

Steps 1, 2 and 3 could be considered the Snowball Sampling method, with the small differentiation that step 2 would add vertices to the sampled database instead of the selection one. The reason for the existence of the selection database is the need for a separate database where we would filter the crawled vertices according to our desired distribution. On the other hand, steps 9 and 10 exist in every sampling method. Thus,

steps 1, 2, 3, 9 and 10 are the common sampling steps and can be arranged to fit any existing sampling method.

The second part of our algorithm consists of our proposed enhancements. Steps 4 to 8 describe the selective sampling enhancement, whereas step 11 is the enhanced subgraph enhancement. As it can be seen, both these improvements are not dependent on the sampling process, but they can be plugged-into any existing sampling method. As mentioned, extensive tests were conducted only with the random sampling method.

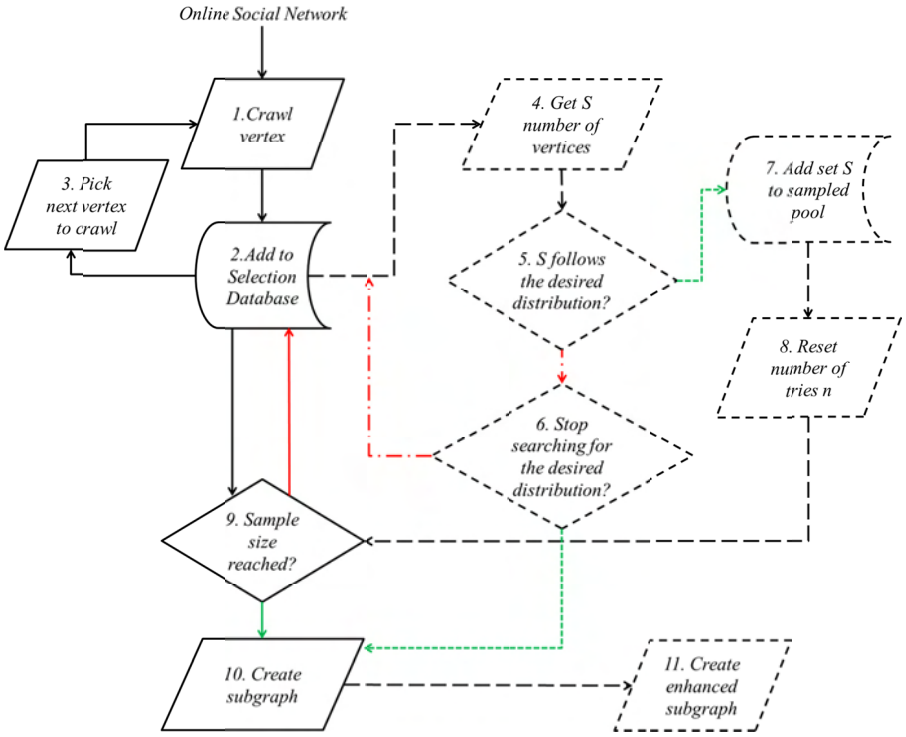

***Figure 10**. Real time enhanced random sampling algorithm*

2.3.3.1. Description

In this section we will present each separate step of the algorithm in detail.

Step *1* - *"Crawl vertex"*, an API request is performed to analyse the in and out going edges of the vertex. When this is applied to an OSN, a user is crawled, leading to the discovery of its neighbouring vertices.

In step *2* - "*Add to Selection Database*", the crawled vertex is added to that temporary database. Again, when we apply this procedure in an OSN, the user's neighbours, both in and out, are stored in this database. For every vertex added to the Selection Database, we check through step *9* if the requested graph size has been reached. When the required size is reached, the normal Random Sampling method continues to step *10*.

In step *3* - *"Pick next vertex to crawl"*, the next vertex to be crawled is picked. In the early iterations of the sampling process, the next vertex is picked amongst the discovered neighbouring vertices of the vertex crawled in step *1*. This means, that during the initial phases of the crawling process it is fair to use more than one API request to discover a large enough set of vertices. However, in later iterations, since we already have discovered a large number of vertices, there is no need to waste requests here. We should note, that step *3* differentiates based on the sampling method used. For example, in the Breadth First crawling we would crawl every first level neighbour of the crawled vertex, while for Random crawling we would randomly choose one.

In the following, step. *4* - *"Get S number of vertices"* we randomly pick a set *S* of vertices from the Selection Database. The number of vertices within *S* could be predefined in a pure number form or in a percentage value of the total vertices in the Selection Database. During our experimentation, we observed that for very small set sizes, the sampled pool of step *7* would fill in a very slow pace; so slow that it would impede our real time sampling process. Furthermore, we decided to determine the size of *S* based on the size of the required sample. We started with a percent value of the number

of vertices our desired sampled graph would have. Initially we tested values below 1%, which would be ideal for the constant monitoring of the sampled pool. Unfortunately, the pace of the sampled pool fill was still too slow. So we incrementally increased the percentage, until a relatively fast fill rate of our pool was found, while at the same time the ability to monitor data was preserved. Upon reaching the *S=2%* value, a respectable pace was achieved and the graph data could be easily monitored. In addition, when we tried to test higher values, we noticed an increased rejection rate from the distribution check step *5*.

In step *5* - *"Distribution Check"*, we check whether the set *S* of vertices taken from step *2* – *"Selection Database"* follows the desired distribution. As we already mentioned, this is decided by the power law percentages we would like to use. In our case, these were 3 distinct power law distributions, *85-15*, *80-20* and *75-25*. The check is performed utilising information crawled during step *1*, the in and out degree of the vertex. If the set of vertices *S* follows the desired distribution, then it is added to the Sampled Pool at step *7*. If it is not, then we move to the control loop at step *6*.

Step *6* – *"Loop Control"*, ensures that the sampling process in not stuck into an infinite search for a specific set *S* of the desired distribution. It is entirely possible that the distribution we request might not even exist within the crawled OSN graph. Furthermore, it prevents the algorithm from looping on occasions where the distribution existed in the graph and the sampled pool had not reached the requested size. The number of tries to discover a set *S* with the desired distribution is *n=10*. Step *8* resets that number after every successful addition of a set *S* into the Sampled Pool.

In the step *7* - *'Sampled Pool"*, every set S that follows the desired distribution is added to this pool. Upon ending the sampling process without the required number of

vertices in the Sampled Pool, the sampled graph is created with whatever number of vertices available.

As described above, the only purpose of Step *8 – "Reset n"* is to reset the number of tries available for the discovery of a set *S* with the desired distribution. It does not matter if the addition happens in the 1st or in the 9th iteration of step *6* upon of a successful finding of a set *S*, the number of tries is reset to 10. Once again, we ended up picking our number of tries based on experimentation. Our aim was efficiency and speed in discovering a set *S*. When we used *n>>10* the sampling process was stuck to a finite but really long loop and when *n<10* even in the most probable distribution, no sets *S* were found and the Sampled Pool stayed unpopulated.

At step *9* - *"Sample Size Check"*, we monitor the sample's size. Usually the size is set in number of vertices or edges. In our tests, since we knew exactly how big the TGs were we used percentage values. Thus, we started sampling until we reached the 10% of vertices included in the TG and continued with 10% increments up to 90%. Realistically, large sample sizes, are not feasible in modern OSNs. We only included them in order to compare the efficiency of our sampling enhancements in greater sample sizes. When the required sample size is reached, we move to step *10*.

In step *10* - *"Create Subgraph"*, we use the vertices from the Sampled Pool to create the subgraph. Note that in case that Selective Sampling enhancement is not used the subgraph is created from the Selection Database with the desired sample size. All random sampling method (such as Random Node or Edge sampling, Random Jump sampling, etc.) end in this step.

In this last step *11* - *"Create Enhanced Subgraph"*, vertices with edges adjacent - but not included- to the vertices in the Sampled Pool or the Selection Database are added to the subgraph. These extra vertices and their edge information is available to us, via

requests made in step *2*. This step is the second proposed enhancement, named "Enhanced Subgraph" and is independent of the sampling method or of the use of the Selective Sampling enhancement. When only the normal Sampling method is performed, the vertices are picked from the Selection Database. When we use the Selective Sampling method though, the vertices are picked from the Sampled Pool.

We expected both enhancements to improve the sampling process. For the Selective Sampling we thought that, a sample of the correct distribution would be of higher quality. As for the Enhanced Subgraph procedure our concept was that more information would not hurt the portraying properties. In the end, both enhancements offered some improvement over the normal sampling method. The Enhanced Subgraph gave a fast and reliable way to increase the accuracy of the subgraph properties, while the Selective Sampling improved a slightly different area, namely the resource management.

2.3.4. DATASET

In order to be able to analyse or implement any changes and improvements to sampling methods, we aimed acquire up to date Twitter graph samples as test data. So we proceeded with the sampling process shaped by the recent changes in the API policies of Twitter. For the interest of speed, we chose 30 seed vertices, the top 30 Greek users based on the number of followers (ingoing edges), and we performed a Breadth First Sampling/Crawling on each of them. This allowed us to perform 10,500 requests per hour.

During the crawling period, every seed had at least every second level neighbour crawled. In some cases, the crawling process reached up to the fourth level neighbours. Since our goal was to perform sampling tests in different datasets, we decided to keep data separated based on the seed vertex. For the occasions where the acquired graph's data (such as the vertex and the edge count) was too small, we decided to merge some

crawled datasets in order to could reach a sufficient graph size. The crawling period lasted three weeks and was performed in March 2012.

More than 200000 twitter users and their connections were analysed. Upon completing the crawling process, about 60GB of Twitter graph data were obtained. Containing more than 93 million vertices and 570 million edges. Because of the multiple seed sampling procedure, some overlapping was inevitable. From the total of 30 samples, we merged the smallest 20 into 4 large graphs, so that each graph would contain at least four million vertices.

2.3.5. ANALYSIS TOOLS AND EVALUATION

Speed was the only factor when preparing our analysis approach and tools for our experimentation. To achieve the desired speed, we had to select a tool that could quickly handle huge graphs. After testing most graph analysis tools, which are mainly focused on visualization, we ended up using R [106] and igraph [107]. R is a multipurpose programming environment that is supported by a large number of independently developed packages. One of these packages is igraph, which is heavily focused in network and graph analysis. Igraph provides many predefined algorithmic procedures and is fairly easy to use.

We compiled 5 different sampling methods and 6 graph property extraction and visualization algorithms in R format. Subsequently, we used igraph built-in functions to test them in our test graphs. This approach was quite fast but utilised RAM heavily. Due to these RAM requirements, we adopted the use of cloud computing, also seen in [97]. We ended up parallelizing our analysis process, utilizing 18 to 36 GBs of RAM per TG. This was possible by employing the infrastructure of high memory instances provided by

Amazon Web Services[29]. The experimentation and debugging processes lasted less than 2 months, while the analysis of every method in all TGs was done within a month.

The evaluation process of the enhanced sampling methods, was based on several graph properties. Comparing our TGs' properties with the sampled properties of each method, we calculated the mean percentage error of each property and the overall mean percentage error. Percentage error is defined, as the relative error between estimation and exact value as a percentage. It is given by the following formula:

$$\delta = \frac{|x - a|}{a} \times \frac{100}{100} = \frac{|x - a|}{a} \times 100\%$$

where δ is the percentage error symbol, x is the estimation and a is the exact value. In our testing and evaluation, x is the value of the property of a TG's sample and a is the exact value of the whole TG. For example, the mean value of the clustering coefficient of a 10% sample of the first TG would be x, while the exact clustering coefficient for the whole first TG would be a. Percentage error values close to zero, indicate a perfect match between the TG sample property value and the TG property value. In essence, percentage error gives a perception of accuracy.

The graph properties we considered in our evaluation were the Number of Edges, the Mean Degree, the Clustering Coefficient, the Assortativity, the Number of Components (NoC) and the time required for each sampling process. In detail, we compare the number of edges acquired by each sampling process, the mean degree of its vertices, the global Clustering Coefficient, the global assortativity and the number of components that constituted our sample. Additionally, we also used edge discovery along with the aforementioned properties in order to include cases where sampling is performed not only for graph analysis, but also for graph drawing purposes. The time required was

[29] http://aws.amazon.com/, Last Retrieved: 13/01/2016

added so we could highlight the time difference in relation to the method used and the sample size needed.

2.3.6. EXPERIMENTATION

As mentioned, we crawled Twitter in order to create a set of graphs on which we would conduct the testing of our sampling enhancements. We started with 30 seed vertices, to get 30 graphs. Only 10 of them were in the order of million edges. So, we merged the 20 remaining into 4 large graphs, with an edge number close to the mean of the 10 largest graphs. The 14 graphs we ended up with are the previously explained TGs.

The mean values of all TGs along with their minimum and maximum values are presented in ***Table 1***. It is evident that all the TGs are pretty dissortative with very low clustering coefficients. Meaning that the graph is loosely connected and uneven, in the sense that vertices of similar degree are rarely connected with each other. Every TG has a single component. The single component attribute, is a direct result of the sampling method used to obtain our test graphs, which was Breadth First Sampling.

Table 1. *Test graphs mean and marginal values*

	Edges	**Vertices**	**Mean Degree**	**Clustering Coef.**	**Assortativity**
Max	69872473	8444642	19.61	0.022451	-0.31064
Mean	40731274	6643998	12.14	0.013787	-0.44001
Min	25658132	4662987	9.74	0.007862	-0.51234

We tested 8 variations of the Random Node Sampling in these 14 TGs. These variations were:

1 *Random Node Sampling*, the most common sampling algorithm without any modification, referred as RNS.

2 *Random Node Sampling with Enhanced Subgraph enhancement*, as described in *Section 2.3.3* with step *11*, referred as RNSE.

3 *RNS with Selective Distribution enhancement of 85-15*, the RNS method with the addition of our proposed Selective Distribution enhancement, with a requested distribution of *85-15*, as described in *Section 2.3.3* in steps *4 to 8*, referred as RNS 85-15.

4 *RNSE with Selective Distribution enhancement of 85-15*, the RNSE method with the addition of our proposed Selective Distribution enhancement, with a requested distribution of *85-15* as described in *Section 3.3* in steps *4 to 8*, referred as RNSE 85-15.

5 *RNS 80-20*, see variation *2*, with a requested distribution of *80-20*.

6 *RNSE 80-20*, see variation *3*, with a requested distribution of *80-20*.

7 *RNS 75-25*, see variation *2*, with a requested distribution of *75-25*.

8 *RNSE 75-25*, see variation *3*, with a requested distribution of *75-25*.

Each of the aforementioned variations was tested multiple times in every TG, as the same sampling procedure was conducted in 10 iterations. For each iteration, we sampled the TG in 9 different percentages. Despite the fact that in a real OSN sampling situation, sampling size targets can only be numerically set (based on either the vertices or the edges), we decided to take that approach because of the varied size of our TGs, so that we can informingly compare the results and later calculate the mean values, as well as the sampling error for every TG.

In the following section, only the mean relative percentage error of the sampled graph is presented, with respect to the initial TG property values. It should be noted that the variance for each iteration of every TG was very small. In total, the algorithm was applied 10080 times, with a total calculation time of approximately 800 hours. When

compared to [99], where the authors required approximately 510 hours in order to analyse four algorithms in 4 separate graphs of various OSNs with a mean size comparable to our smallest TG, our analysis was completed in an extremely quick fashion. The mean sampling time for every iteration was approximately 286 seconds. Of course, sampling time for really small percentages was much lower than the average, while in larger sample sizes sampling time was several times higher than the average.

2.3.7. RESULTS

Percentage error of each property (Calculation Time, Vertices, Edges, Mean Degree, Clustering Coefficient, Assortativity, Components) was averaged in a single mean percent value. This allowed us to preview through a single numerical value the overall accuracy of our proposed sampling method, always in comparison to the exact values of the TG's properties. We will present in detailed error charts and tables the mean sampling accuracy of every method for every sample size of every TG.

In every chart and table of this section, we present the results in percentage error values. Every percentage error value is calculated as follows. We started with a single sampling method iteration of a single sampling percentage size in the first TG, for example with RNS of 10% sample size in the first TG. Then, we calculated the exact values of the following properties, Number of Edges, Mean Degree, Clustering Coefficient, Assortativity and Number of Component, as seen in the top greyed line of ***Table 2***. We then performed 9 more iterations of the same sampling method as seen between the greyed lines of ***Table 2***, with the same sample size, in the same TG. Then we averaged the values from these 10 iterations in a single line, as in the bottom greyed line of ***Table 2***.

Table 2*. Every iteration of RNS with 10% sample size in the first TG*

	Calculation Time	*Vertices*	*Edges*	*Mean Degree*	*Clustering Coefficient*	*Assortativity*	*Components*
Sample Size: 10.00%	2.91	828940	614735	1.4832	0.0179622	-0.3153210	648359
	2.84	828316	584313	1.4108	0.0142348	-0.3538748	651912
	2.84	828513	596649	1.4403	0.0188440	-0.3235506	657758
	2.87	828985	621502	1.4994	0.0188810	-0.2754536	647788
	2.87	828842	612861	1.4788	0.0190295	-0.3601920	650451
	2.88	828418	615159	1.4851	0.0198704	-0.2915701	649422
	2.88	828742	600564	1.4493	0.0197270	-0.2849092	655273
	2.88	828840	603931	1.4573	0.0176338	-0.3253891	647863
	2.85	828622	604050	1.4580	0.0182787	-0.3069020	651455
	2.83	828500	579373	1.3986	0.0197688	-0.2989926	657919
	2.86	**828672**	**603314**	**1.46**	**0.0184230**	**-0.3136155**	**651820**

***Table 3**. Random Node Sampling results for the first TG*

Size	*Time*	*Vertices*	*Edges*	*Mean Degree*	*Clustering Coefficient*	*Assortativity*	*Components*
10%	2.86	828671.8	603313.7	1.46	0.01842	-0.31362	651820.0
20%	4.37	1542028.3	2106751.6	2.73	0.01786	-0.31186	990795.5
30%	6.10	2155321.3	4052288.7	3.76	0.01741	-0.32069	1159386.0
40%	8.12	2684508.3	6373124.3	4.75	0.01746	-0.31013	1185728.0
50%	10.20	3138528.7	8655063.9	5.52	0.01765	-0.31013	1191189.7
60%	16.86	4157132.2	15126071.5	7.28	0.01737	-0.31274	951865.6
70%	19.66	4481967.3	17600094.0	7.85	0.01792	-0.31158	866541.8
80%	28.71	5143861.5	23132987.7	8.99	0.01762	-0.30764	517844.7
90%	39.33	5548756.7	27013947.8	9.74	0.01777	-0.31047	276971.0
100%	**-**	**5948844.0**	**31040820**	**10.44**	**0.01773**	**-0.31064**	**1.0**

Each averaged value was entered in a table (similar to ***Table 3)*** for a particular sampling method and a TG. The line on the top of ***Table 3*** is the average value from ***Table 2***. The exact TG property values are presented in the bottom line of ***Table 3***. The percentage error values were then calculated, which resulted in ***Table 4***. In the last two columns of ***Table 4*** we present the mean values of every line. Specifically, in the column named “Mean Values” we only take into account the Mean Degree, the Clustering Coefficient, the Assortativity and the Components percentage errors. In the column named “Mean Values ED”, we averaged all the percentage errors used in “Mean Values”

column, with the addition of the Edges percentage error. As we mentioned in a previous section ED stands for Edge Discovery and was used to evaluate our sampling enhancements in cases where edge discovery matters.

These steps were repeated for every each of the 14 TG, and from each of the 14 respective tables we averaged the “Mean Value” and the “Mean Values ED” percentages. These two percentages, are presented in the accuracy tables. The closer a value is to zero, the greater the accuracy. The accuracy table for the RNS method can be seen in ***Table 5***. The “RNS” column presents the percentage error values for every sample size, using the Random Node Sampling method. The “RNSE” column has the percentage error values of Random Node Sampling with the Enhanced Subgraph enhancement. The last two columns, “RNS ED” and “RNSE ED”, present the percentage error values when we consider the “Mean Values ED” for the RNS and the RNSE methods respectively. With green shading we denote the best values of each pair of the RNS and the RNSE values. These values are also presented in the form of a chart in ***Figure 11***.

Table 4*. Percentage error of RNS in the first TG*

Sample Size	*Vertices*	*Edges*	*Mean Degree*	*Clustering Coefficient*	*Assorta -tivity*	Components	*Mean Values*	*Mean Values ED*
10%	86.070%	98.056%	86.047%	3.902%	0.956%	65181900 %	30.302%	47.24%
20%	74.079%	93.213%	73.817%	0.699%	0.392%	99079450 %	24.969%	42.03%
30%	63.769%	86.945%	63.968%	1.802%	3.233%	115938500 %	23.001%	38.99%
40%	54.873%	79.469%	54.503%	1.550%	0.166%	118572700 %	18.739%	33.92%
50%	47.241%	72.117%	47.150%	0.479%	0.164%	119118870 %	15.931%	29.98%
60%	30.119%	51.270%	30.268%	2.047%	0.676%	95186460 %	10.997%	21.07%
70%	24.658%	43.300%	24.743%	1.052%	0.303%	86654080 %	8.699%	17.35%
80%	13.532%	25.476%	13.813%	0.621%	0.967%	51784370 %	5.134%	10.22%
90%	6.725%	12.973%	6.698%	0.234%	0.057%	27697000 %	2.330%	4.99%

Table 5*. Mean percentage error*

Method / Sample size	RNS	RNSE	RNS ED	RNSE ED
10%	**29.85%**	41.93%	**46.91%**	49.96%
20%	**25.45%**	29.48%	42.41%	**35.84%**
30%	21.81%	**21.21%**	38.08%	**26.07%**
40%	18.77%	**15.19%**	34.00%	**18.93%**
50%	16.28%	**11.07%**	30.26%	**13.89%**
60%	10.46%	**4.35%**	20.64%	**5.54%**
70%	8.51%	**2.90%**	17.18%	**3.70%**
80%	4.81%	**0.86%**	9.93%	**1.11%**
90%	2.37%	**0.20%**	5.02%	**0.26%**

2.3.7.1. RNS with Enhanced Subgraph Enhancement

Remarkably, throughout testing, RNS –with or without any enhancements- accurately portrayed the clustering coefficient and assortativity values in any sample size. On the other hand, it completely failed in portraying the mean degree and component number values. Nonetheless, since our TGs are dissortative and with a low clustering coefficient, we cannot conclusively associate RNS with the abovementioned properties. Furthermore, RNSE maintained a rather stable mean degree and component size projection in every sampling size. Despite that, the clustering coefficient and assortativity values were far off their real values in samples sizes lower than 20%. Above the 20% point RNSE outperformed RNS in every property. In the edge discovery evaluation, above the 20% sample size, accuracy is heavily favoured from RNSE. Our tests showed an improvement of RNSE over RNS for Edge Discovery in sample sizes larger than 40%. The set values can be seen in ***Table 5*** and also in chart form in ***Figure 11***.

As mentioned, we noticed that on every test RNS depicted the clustering coefficient and assortativity values throughout the different sample sizes. Contrary, it failed completely on the mean degree and the component size. Since our initial graphs

have already a low clustering coefficient and low assortativity, we cannot conclusively relate this with our data and/or the sampling method. While on the one hand RNSE maintained a fairly stable mean degree and component size for different sample sizes, the clustering coefficient and assortativity were marginally improved on sample sizes greater than 20%. Furthermore, the edge discovery evaluation is heavily favoured by RNSE, with a mean accuracy improvement of 40% over RNS.

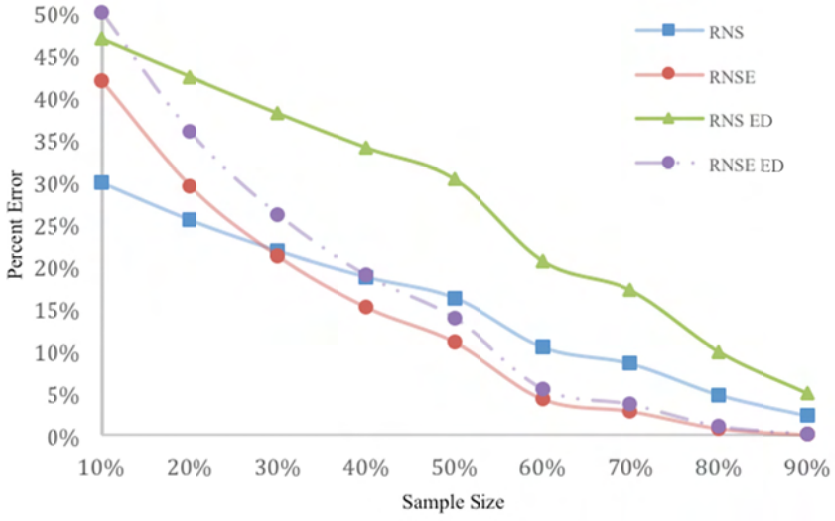

Figure 11. *Mean percentage error of RNS*

In ***Figure 12***, we present the improvement of the RNS mean percentage error over time. We can observe, that although both RNSE and RNSE ED have better overall accuracy in sample sizes over 20%, both of them need more time to produce these results. Nonetheless, we should note that this should not hamper the ability for real time sampling, since the enhanced subgraph improvement is applied at the end of the sampling process. It will, however, burden the total time required to create the graph. As such, time monitoring proves a useful addition, towards the precise evaluation of the RNS, and the RNS with the Enhanced Subgraph method enhancements.

Hence our sampling results can be interpreted based on the sampling goals and needs of each researcher. When the aim is to portray an OSN as accurately as possible, with accuracy based on property values, RNSE is way more accurate than simple RNS. However, this gap in accuracy is only noticeable in larger sample sizes, from 20% and above. In real time OSN sampling scenarios where we can't exactly determine the size or the percentage of our sampled graph over the whole OSN, we strongly recommend the extensive sampling of the desired OSN in order to approach the 20% bound. In order to end up with an important –size wise– sample of the requested OSN. Which in turn, would enable the efficient utilization of RNSE method.

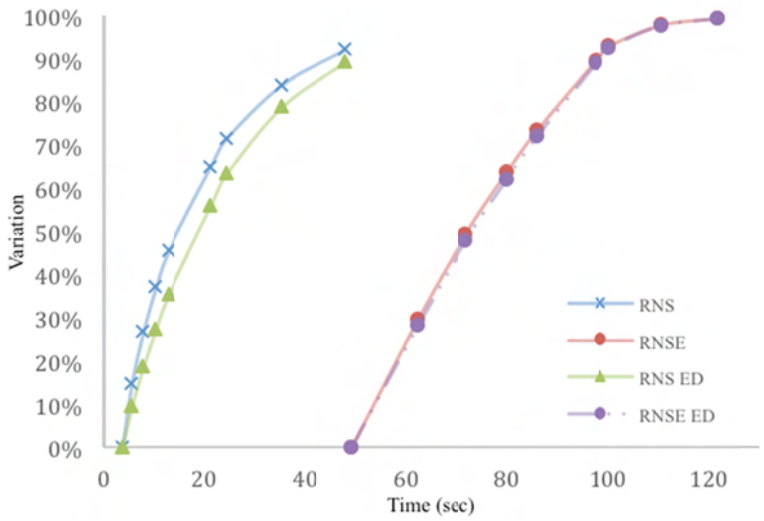

Figure 12*. Improvement of mean percentage error over time for RNS*

In cases where edge discovery is an important factor of the sampling process the RNSE is recommended for every sampling size. Even though, for 10% sample sizes the RNSE is slightly less accurate for Edge Discovery situations, the difference is so small that this slight loss of accuracy is negligible, when compared to the improvement seen in

samples over 10%. Once more, the accuracy of RNSE ED over RNS ED is growing exponentially as the sample size increases.

2.3.7.2. RNS with Selective Sampling Enhancement

In the RNS with Selective Sampling Enhancement, the results were mixed. Out of all the three distributions (75-25, 80-20 and 85-25) only the 85-15 underperformed significantly, compared to the RNS and the RNSE. Both the 80-20 and the 75-25 selective sampling methods provided similar –if not a bit worse– results, with those of the RNS method. However, by using the Selective Sampling enhancement we managed to obtain the aforementioned results with a sufficiently smaller number of vertices. But how can this be meaningful in an OSN sampling scenario?

***Table 6**. Mean percentage error of RNS 85-15*

Evaluated Method / Sample size	RNS	RNSE	RNS ED	RNSE ED
10%	**33.49%**	46.40%	**49.85%**	54.97%
20%	**29.56%**	35.52%	46.13%	**42.68%**
30%	**25.89%**	27.54%	42.13%	**33.39%**
40%	23.68%	**22.23%**	39.09%	**26.93%**
50%	20.50%	**17.03%**	34.77%	**20.65%**
60%	19.12%	**14.78%**	32.33%	**17.93%**
70%	17.06%	**13.83%**	29.21%	**16.66%**
80%	15.85%	**12.32%**	26.51%	**14.78%**
90%	14.82%	**12.34%**	24.43%	**14.78%**

First and foremost, as we mentioned in *Section 2.3.1*, OSNs frequently employ restrictions the access of their data. Thus, not only time and computational resources, but also access resources should be considered when designing and executing a sampling process in such OSNs. A smaller number of vertices required means a conservation in these access resources, which usually are time-based. Thus, that conservation in time

sensitive resources, results in less time required for the sampling process. But before we analyse the exact losses and gains of the proposed selective sampling, we provide the mean percentage error values.

In the 85-15 selective sampling, apart from the underperformance we observed that Enhanced Subgraph Enhancement reached its peak accuracy at the 80% sample size, on both the evaluation methods. The precise values can be found in ***Table 6*** and the respective graph can be seen in ***Figure 13***. Essentially after the 80% point it was impossible to further gather any sets following the required distribution, because the graphs degree distribution is obviously lower than 85-15. As for the results, the RNSE 85-15 fares better than the RNS 85-15 on larger samples, similar to RNS. However, the tipping point in the particular method is 30% sample size, where edge discovery is not considered and the difference is marginal. On the other hand, when edge discovery is added to the evaluation process, RNS ED 85-15 is only outperforming RNSE ED 85-15 in sample sizes less than 10%, but underperforms heavily in every other occasion. In ***Figure 13*** we can clearly observe the lack of accuracy of RNS ED 85-15 compared to RNSE ED 85-15. Time wise, Enhanced Subgraph improvement did not tax the sampling process. But, it provided greatly improved results in the same processing time, as seen in ***Figure 14***.

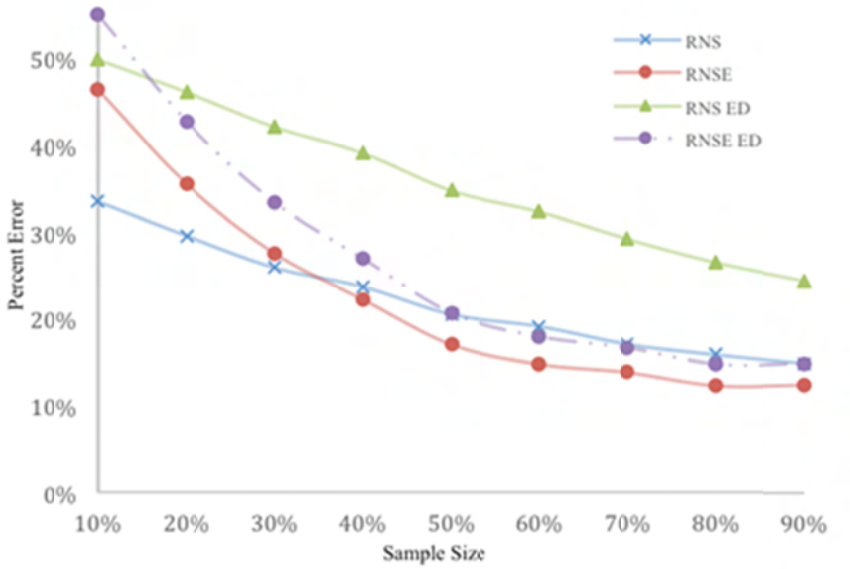

***Figure 13**. Mean percentage error of RNS 85-15*

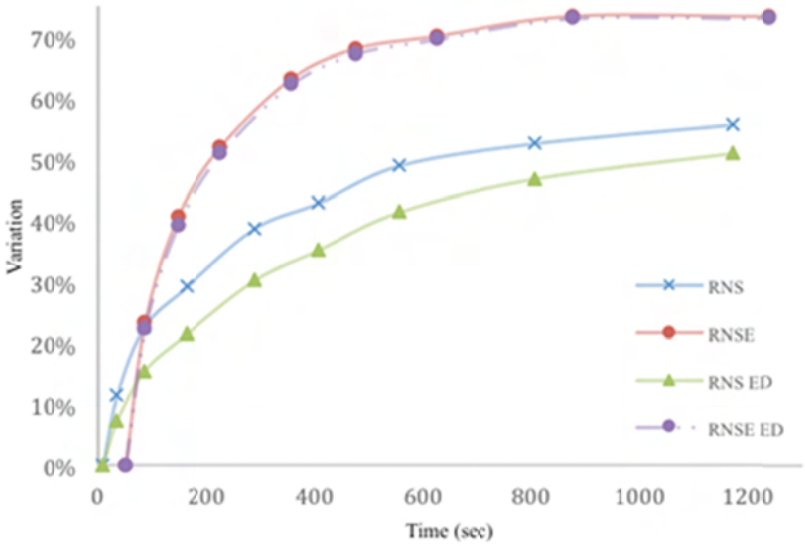

***Figure 14**. Improvement of mean percentage error over time for RNS 85-15*

The results of the 80-20 and the75-25 Selective Sampling Distributions were similar. That was due to the fact that any set *S* of vertices that follows the 75-25 distribution also follows, by definition, the 80-20 distribution. The RNS without the

enhanced subgraph enhancement for both the 80-20 and the 75-25 only offers an increased accuracy in sample sizes lower than 30%. Their detailed values of both distributions can be observed in ***Table 7***. With the enhanced subgraph applied and the sample size increased over that 30%, the accuracy improvement is apparent in both distributions.

This is more profound in the edge discovery evaluation, where the accuracy improvement for both distributions is evident from the 20% sample size and onwards. Nonetheless, the great difference in the accuracy of RNS ED and RNSE ED can be seen in sample sizes greater than 30%. In ***Figure 15*** and ***Figure* 16** we can clearly observe the point in which RNSE ED intersects with RNS ED along with the respective accuracy difference above the 30% sample sizes.

Table 7*. Mean percentage error of RNS 80-20 and RNS 75-25*

RNS 80-20	RNSE 80-20	RNS ED 80-20	RNSE ED 80-20	Evaluated Method / Sample size	RNS 75-25	RNSE 75-25	RNS ED 75-25	RNSE ED 75-25
31.48%	46.80%	**48.36%**	55.34%	10%	**31.75%**	46.68%	**48.56%**	55.23%
27.52%	35.07%	44.63%	**42.27%**	20%	**27.49%**	35.20%	44.61%	**42.36%**
24.12%	25.92%	40.83%	**31.67%**	30%	**24.34%**	25.96%	41.00%	**31.71%**
20.62%	**18.50%**	36.45%	**22.85%**	40%	20.69%	**18.54%**	36.51%	**22.89%**
17.11%	**12.46%**	31.56%	**15.58%**	50%	17.20%	**12.45%**	31.64%	**15.58%**
13.70%	**7.76%**	26.23%	**9.79%**	60%	13.68%	**7.83%**	26.24%	**9.86%**
10.19%	**4.20%**	20.33%	**5.37%**	70%	10.25%	**4.32%**	20.43%	**5.49%**
6.88%	**1.79%**	14.05%	**2.32%**	80%	7.00%	**1.86%**	14.25%	**2.40%**

3.45%	**0.38%**	7.09%	**0.51%**	90%	3.46%	**0.44%**	7.32%	**0.58%**

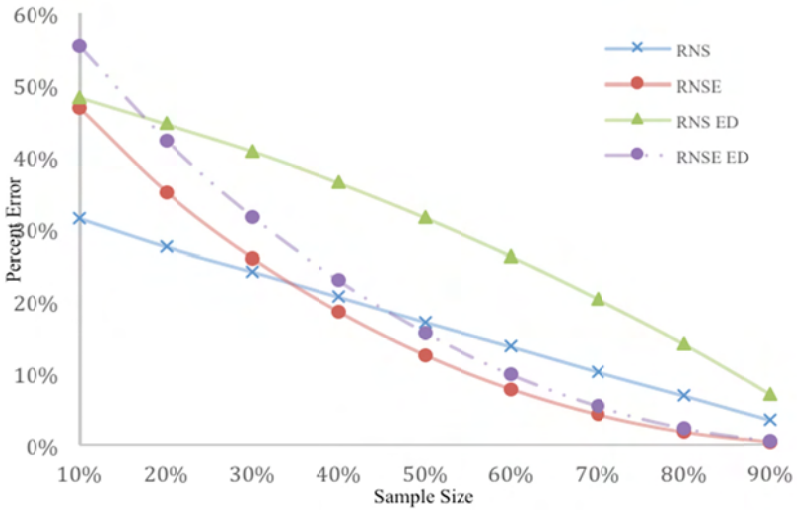

***Figure 15**. Mean percentage error of RNS 80-20*

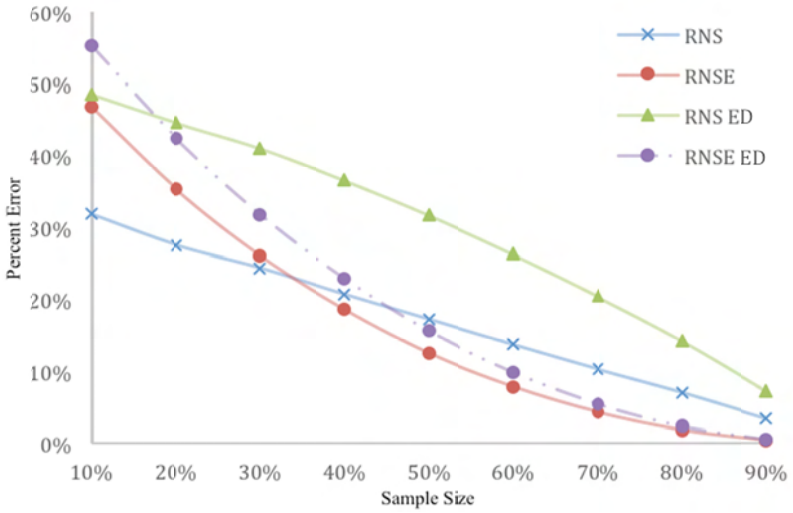

***Figure 16**. Mean percentage error of RNS 75-25*

Regarding the required time, we can observe the improvement of the results over time. In ***Figure 17***, the chart of sampling with the 80-20 selective is perceivably identical with the chart seen in ***Figure 18*** of the 75-25 distribution. The marginal difference can

only be noted by the values depicted in the previously provided **Table 7**. As is the 85-15 selective distribution, time required is not affected by the usage of Enhanced Subgraph enhancement.

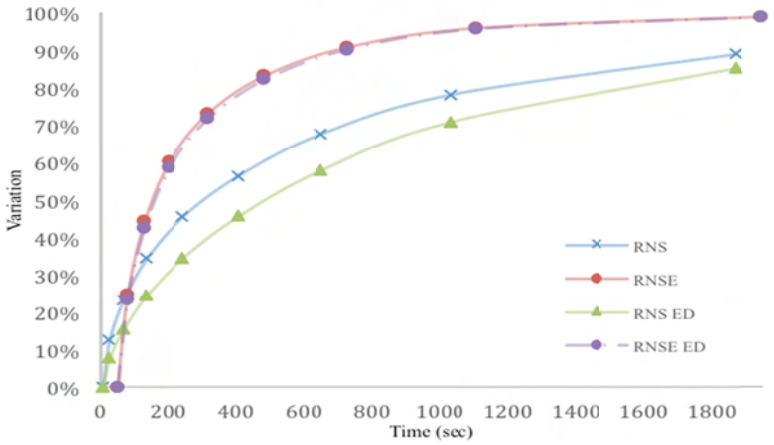

***Figure 17**. Improvement of mean percentage error over time for 80-20*

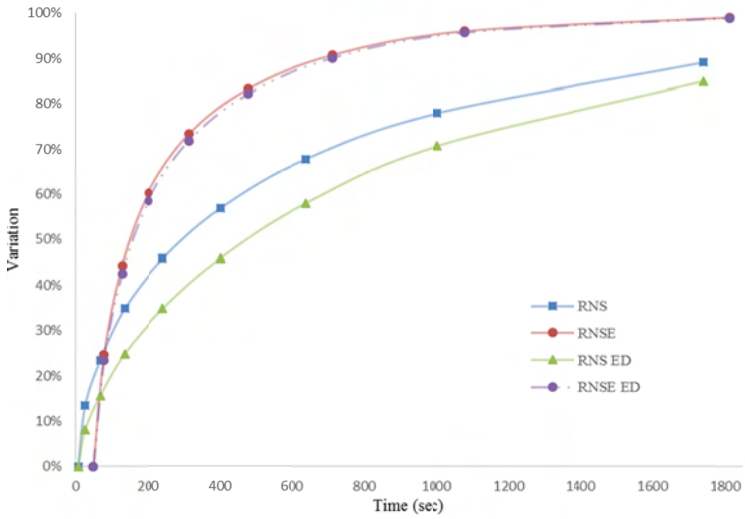

***Figure 18**. Improvement of mean percentage error over time for 75-25*

Therefore, the most successful selective distribution method, based on percentage error over every sample size is the 80-20. An even finer tuning could possibly reveal

another distribution (e.g. 79-81) as the most fitting distribution. But as we saw from the 75-25 and 80-20 comparison, where the difference in the required set was small, the differences would be even smaller. If we compare the results from selective sampling with 80-20 distribution and RNS, as seen in ***Table 8***, we see that the RNS provided universally better results. The mean inaccuracy of the RNS 80-20 Selective Sampling in relation to the RNS, was almost 9.5% percent. The respective inaccuracy values of the RNS 85-15 and 75-25, are shown in

Samplin g Method	Mean Inaccuracy
RNS 85 - 15	250.44%
RNS 80 - 20	**9.55%**
RNS 75 - 25	12.47%

. Apparently, RNS 85-25 should be discarded as extremely inaccurate, whereas RNS 80-20 is by a small margin the most accurate out of the three distributions.

Table 8*. RNS and RNS 80-20 comparison, with or without Enhanced subgraph enhancement*

RNS	RNS 80-20	RNSE	RNSE 80-20	Evaluated Method / Sample size	RNS ED	RNS ED 80-20	RNSE ED	RNSE ED 80-20
29.85%	31.48%	**41.93%**	46.80%	10%	**46.91%**	48.36%	**49.96%**	55.34%
25.45%	27.52%	**29.48%**	35.07%	20%	**42.41%**	44.63%	**35.84%**	42.27%
21.81%	24.12%	**21.21%**	25.92%	30%	**38.08%**	40.83%	**26.07%**	31.67%
18.77%	20.62%	**15.19%**	18.50%	40%	**34.00%**	36.45%	**18.93%**	22.85%
16.28%	17.11%	**11.07%**	12.46%	50%	**30.26%**	31.56%	**13.89%**	15.58%
10.46%	13.70%	**4.35%**	7.76%	60%	**20.64%**	26.23%	**5.54%**	9.79%

8.51%	10.19%	**2.90%**	4.20%	70%	**17.18%**	20.33%	**3.70%**	5.37%
4.81%	6.88%	**0.86%**	1.79%	80%	**9.93%**	14.05%	**1.11%**	2.32%
2.37%	3.45%	**0.20%**	0.38%	90%	**5.02%**	7.09%	**0.26%**	0.51%

Why should the Selective Sampling enhancement be utilised since it is less accurate than RNS? The answer, as mentioned, is resource conservation. During the implementation of the Selective Sampling improvement, we noticed that, apart from the aforementioned inaccuracy, the number of vertices used in each iteration of Selective Sampling was considerably less than of those used in the RNS. The exact number and percent difference of the vertices used in Selective Sampling, in relation to the RNS are presented in ***Table 10***. Despite the low number of vertices used in the RNS 85-15, the inaccuracy of the method is so high that this improvement is a useless trait. On the other side, the RNS 80-20, which provided the most accurate results, presented the lowest conservation value. Specifically, it managed to use a mean of 331473 less vertices to create each subgraph.

Table 9*. Mean inaccuracy of Selective Sampling RNS*

Sampling Method	Mean Inaccuracy
RNS 85 - 15	250.44%
RNS 80 - 20	**9.55%**
RNS 75 - 25	12.47%

As mentioned previously, one request in the Twitter API is needed for the analysis of a vertex degree (in and out) or for the discovery of 5K neighbours (in or out). The least beneficial case, i.e. when each request would have discovered 5.000 neighbours, contrariwise the best case would have been if every conserved vertex was the result of one request call. The respective number of conserved requests could amount from as low as 66 to a maximum of 333473. The time benefit would in turn range from

less than half an hour worth of requests to almost 930 hours. In the posterior case which is the most optimistic, it would mean that 930 less hours are required to attain the same sample with Selective Sampling, compared to RNS. In a moderate assessment, where only 10000 less requests are used, the sampling process would take at least 27 hours less. Thus, we consider that the scalability of Selective Sampling will be invaluable, since for larger samples this reduction in vertices would result in less time dedicated to crawling/sampling. For our total sampling process, even with the moderate time gain of 25 hours per sample analysed, we could preserve up to 300 hours of calculation and sampling time for the 14 TGs.

***Table 10**. Number and percentage of less vertices used in Selective Sampling*

METHOD	NUMBER OF USED VERTICES	PERCENT DIFFERENCE
85-15	3387917	**23.05%**
80-20	4072877	7.49%
75-25	4071613	8.75%
Non-Selective	4403086	-

Since we accurately defined the costs and gains of selective sampling, we can now answer the question, why should we use Selective Sampling methods?

Every distribution provided different results. The RNS 85-15 sampling method was 250% more inaccurate compared to the RNS and used 23.05% less vertices. The RNS 80-20 was the most accurate and provided the second best recourse conservation, having a mean inaccuracy of 9.55% and using 7.49% less vertices compared to the RNS. Finally, the RNS 75-25 had a mean inaccuracy of 12.47% with an 8.75% mean reduction in the number of vertices used, with respect to the RNS. Obviously, the most successful

selective sampling method, which combined remarkable accuracy and adequate resource conservation was RNS 80-20.

2.3.7.3. Stanford Large Network Dataset Collection

Expanding our evaluation, we studied the potential improvements of RNS enhancements in other non OSN graphs. We used a diverse set of various graphs, the Stanford Large Network Dataset Collection (SLNDC)[30]. Which contains, a multitude of graphs originating from various sources, such as online shops and patent offices. Their properties were vastly different. Some were dissortative, others assortative, most were multicomponent and three of them were single component graphs. We chose eight directed graphs, which we will refer to as Stanford Graphs (SGs). Out of these 8 SGs, 5 were multicomponent. This decision was made, so we could further evaluate the improvement of our enhancements in multicomponent graphs, since every TG was single component. Their detailed properties can be seen in ***Error! Reference source not found.***, where assortative SGs are highlighted with green shading. As we can observe, clustering coefficient is also pretty low, but higher than in the OSN TGs. Unfortunately, no relevant publication existed that proved the existence of a Pareto Distribution within these specific, not of social network, graphs. So we decided not to force the selective distribution of 80-20 in these samples.

Table 11. *Detailed properties of every SG*

Graph	Edges	Vertices	Mean Degree	Clustering Coef.	Assortativity	Components
Amazon	3387388	403394	16.79	0.165622	-0.013843	7
Google	5105039	875713	11.65	0.055230	-0.048881	2746
LiveJournal	68475391	4847571	28.25	0.117915	0.055149	1876
NotreDame	1469679	325729	9.02	0.087673	-0.044035	1
Patents	16518947	3774768	8.75	0.067142	0.167667	3627
Slashdot	828161	77360	21.41	0.024157	-0.066329	1

[30] Stanford Large Network Dataset Collection. http://snap.stanford.edu/data/index.html, Last Retrieved: 13/01/2016

Slashdot_1	870161	82168	21.18	0.024108	-0.066612	1
Wikitalk	5021410	2394385	4.19	0.002192	-0.062385	2555

Instead, we estimated the approximate degree distribution of every graph in this set. Alongside the approximation method, the real time character of our sampling method should also be preserved. This forced us to create a really fast and efficient method, in order to approximate the degree distribution of any given graph. We approached the problem, bearing in mind that any solution should be utilised, in parallel to the sampling process. Thereby we chose to use the initial Random Sampled vertices to estimate the degree distribution. By calculating, the mean degree distribution of 0.1% of every SG over 10 iterations, we were able to obtain a really good estimation of the exact distribution value. It was also a really fast method, as it only required one second (at most) for the execution of all the iterations per SG. The accuracy was also remarkable, considering the total time required. For example, the Amazon distribution estimation was 60-40, a mere 1.2% deviation from the actual distribution. The estimated and actual distribution values of every SG are shown in ***Error! Reference source not found.***.

Table 12*. Estimated and actual distribution of every SG*

Graph	Estimation	Actual
Amazon	60-40	61.2-38.8
Google	70-30	72-28
LiveJournal	80-20	77.7-22.3
NotreDame	82-18	80.3-19.7
Patents	70-30	68.7-31.3
Slashdot	81-19	81.1-18.9
Slashdot_1	80-20	81.1-18.9
Wikitalk	78-22	79.5-20.5

Since we are testing a real time sampling method and its enhancements, this efficient distribution estimation method, provided the basis for selective distribution sampling. When we tested the Selective Sampling enhancement in TGs, we aimed to find out if distributions similar to Pareto would work. Similarly, we used the estimated distribution with the Selective Sampling process along with its surrounding distributions.

For instance, in Amazon SG we tested the 65-35, 60-40 and 55-45 distributions. For each SG we found the most efficient distribution out of the triads we tested. We named it Best Distribution (BD). The exact value of BD is irrelevant, since it is defined by the estimation method we mentioned above. A total of five thousand tests were conducted in approximately 60 hours.

To directly compare sampling results, a uniformity of the sampled graphs must exist. For this reason, we split the SGs into three distinct groups based on component size and assortativity. For every Selective Sampling method conducted on these SGs, the only distribution presented is BD. The sampling values for multicomponent graphs can be seen in ***Table 13***.

Table 13*. Mean percentage error for multi component graphs*

Method / Sample Size	Assortative Graphs				Dissortative Graphs			
	RNS	RNSE	RNS BD	RNSE BD	RNS	RNSE	RNS BD	RNSE BD
10%	**34.64%**	85.62%	**42.78%**	98.20%	**67.48%**	99.88%	**73.10%**	112.52%
20%	**27.68%**	56.15%	**31.23%**	70.52%	**49.51%**	52.36%	**67.23%**	69.29%
30%	**24.58%**	36.67%	**32.82%**	45.09%	52.22%	**34.09%**	47.78%	**39.32%**
40%	**23.43%**	23.54%	**20.73%**	32.61%	27.92%	**20.02%**	52.45%	**29.28%**
50%	19.12%	**15.84%**	**18.23%**	21.33%	27.44%	**14.23%**	36.97%	**19.37%**
60%	11.72%	**7.14%**	13.81%	**12.18%**	22.50%	**5.86%**	22.24%	**9.88%**
70%	8.69%	**3.92%**	12.64%	**5.84%**	12.94%	**3.62%**	16.62%	**4.85%**
80%	6.01%	**1.36%**	8.10%	**2.64%**	5.54%	**1.36%**	13.65%	**2.21%**
90%	2.74%	**0.36%**	3.49%	**0.61%**	4.89%	**0.24%**	4.14%	**0.52%**

At first glance, the SLNDC graph sampling results seem consistent with the TGs sampling results. In small sample sizes, RNS is better than RNSE, but slowly loses its efficiency in larger sample sizes. However, the outperformance of RNS over RNSE in smaller sizes is most prominent in assortative graphs, where the tipping point is well after 30% (as it was in TG sampling). This is more evident in RNS BD, which is more accurate than RNSE BD, up until the 60% size. As for dissortative SGs, their sampling results seem to follow the exact patterns we encountered in TGs sampling. Where both

RNS and RNS BD fall behind after 20% sample sizes over RNSE and RNSE BD respectively. More specifically, RNSE underperformed in every property projection compared to RNS, but showed remarkable accuracy in the mean degree. As for the BD pros and cons, for all the SGs, the mean vertices conservation was 5.90% with a mean decline in results of 21.62%. Concerning the mean BD selective sampling conservations, 6.41% less vertices were used to obtain 17.06% worse results, compared to the 80-20 selective sampling results from TGs with respective values of 7.49% and 9.55%.

In the case of single component SGs, the sampling results were almost identical to those of TGs sampling. The sample size over which Enhanced Subgraph sampling becomes more accurate than Random Sampling is 30%. Interestingly, RNS BD compared to RNS, as well RNSE BD compared to RNSE provide similarly accurate results, once more after the 30% sample size. The required time for the RNS sampling processes in these single component graphs is less than 1 second, while the BD sampling methods required less than 10 seconds for up to the 70% sample size and less than 27 seconds for larger sample sizes.

Before we summarize the results of SGs sampling, for both multicomponent and single component dissortative graphs, we should mention that the aforementioned similarities in result improvement of SGs sampling with those of TGs sampling are apparent. In contrast, assortative graphs are not suitable for selective sampling methods and in a sense, neither for enhanced sampling enhancement, since the improvement from both these enhancements is rather insignificant and appears in large sample sizes.

Table 14*. Mean percentage error for single component graphs*

Method / Sample Size	RNS	RNSE	RNS BD	RNSE BD
10%	**39.37%**	54.10%	**43.06%**	65.07%
20%	**33.41%**	35.93%	**33.76%**	40.29%
30%	23.72%	**22.46%**	32.30%	**27.38%**

40%	24.05%	**14.62%**	24.93%	**17.87%**
50%	19.40%	**10.31%**	20.13%	**11.68%**
60%	12.67%	**4.28%**	14.55%	**7.58%**
70%	9.26%	**3.35%**	11.89%	**3.62%**
80%	5.18%	**1.07%**	9.16%	**1.69%**
90%	3.30%	**0.28%**	4.36%	**0.25%**

Thereby, assortativity plays a significant role in the efficiency and the application of RNS and our proposed enhancements. Yet, the most crucial factor in the accuracy of RNS and our enhancements is the component size of each graph, especially for lower sampling sizes. Additionally, on most cases of the SGs ED sampling evaluation, the accuracy on edge discovery attribute was noteworthy.

More specifically, in one out of the three single component SGs, the Edge discovery resulted in almost 3% reduction of the overall accuracy. Whereas on the other two single component SGs, we observed an improvement of the sampling results ranging from 1% to around 36%, when co-evaluating the edge discovery property, Of the five assortative SGs, only one SG benefited from the Enhanced Subgraph enhancement, with accuracy improvement over the 60%. The remaining four SGs, upon utilisation of Enhanced Subgraph sampling, had their mean accuracy diminished from 5% to an astonishing 123%. On multigraph SGs, the only graph benefited by the Enhanced Subgraph enhancement had the lowest mean degree and a negative assortativity. Thus, the sampling accuracy in every assortative graph, was hurt by the Enhanced Subgraph sampling enhancement. Overall, in every multicomponent SG, Enhanced Subgraph favoured the mean degree and component values, but failed in improving assortativity and clustering coefficient projections.

2.3.8. CONCLUSIONS

In *Section 2.3*, we presented a new online sampling approach, which merges sampling and crawling into the same procedure in OSNs environments. Our approach maintained the real-time character of the process and improved its results, while our

evaluation provided several interesting results. Despite the fact that these results were not universally improved over the simple RNS method, improvements could be identified in several other areas, such as in execution time and computational resources.

Every sampling case should be addressed with a different approach. When the requested OSN sample size is small, 20% of the sampled OSN or lower, the most effective sampling method is RNS. Which presents a significant accuracy projection for the clustering coefficient and the assortativity values. In various OSNs sampling cases, where data access is limited and vertex discovery is impeded by the network, each vertex crawled should be considered as a valuable resource. If the aimed OSN is bound by similar restrictions as Twitter and the requested sample size is larger than 20%, RNSE provides the best results without affecting execution time of the sampling process. It also provides a better output of mean degree and component number values, compared to simple RNS.

While in most cases, where we are unable to pre calculate the required sample size and the respected OSN is limiting our data access, the best sampling method is RNSE with the selective distribution enhancement of 80-20. This demonstrates the relative success of our approach, since most of the OSNs sampling situations fall under this category. When applying RNSE 80-20, the loss of accuracy is rather small compared to the speed improvement, over the simple RNS method. A speed improvement which is due to fewer requests utilised as shown in ***Table 10***.

Furthermore, on online sampling cases where power law distribution is not proven, we presented a real time procedure to estimate the degree distribution of the sampling network. With small time-cost, we achieved a great estimation accuracy. In these cases, the selective distribution enhancement is applied together with the estimated

degree distribution. On both OSNs and non-OSNs and with sample sizes more than 70%, RNSE 80-20 and RNSE BD -respectively- provided results almost identical to RNS.

3. MULTI LAYERED INFORMATION DIFFUSION

3.1. Information networks

OSNs are complicated structures, while in the previous sections we discussed their underlying construct, properties and some applications, we should also consider the information propagating in them. Information that is constantly exchanged online among peers, such as friends, acquaintances, family members, colleagues and even unknown individuals. The context varies: from local or world news, to general and scientific facts, quotes, personal preferences and affiliation ties or even recommendations.

Besides the study of such information by itself, researchers have focused also on the direction, speed, depth and several other properties of information propagation. Since 1950, information flows have been in the centre of multidisciplinary scientific researches. Up until internet penetration in the late 90s, these studies were based in traditional offline social networks. Observations of “offline” information flows studies, such as the two-step flow of communication and the importance of weak ties were verified in several “online” studies. For example, a verified study in both offline and online networks, was the two-step theory that presents the micro and micro connections of information flows within social networks. It is observed that diffused information, flows from the initial submission to several users of a particular OSN and then transitions to several other networks. Furthermore, the original content is enhanced or weakened, according to its topic, the dynamics and the exposure of each OSN. In that context, each OSN could be considered as a layer of information flows. These OSNs interact with multiple other similar networks, and they are connected through multiple information flows.

Diffusion is described by sociologists as the spreading of a behaviour or a novel idea. In information science, diffusion is used to describe how information is spread amongst entities. In the premise of OSNs, diffusion describes the spread of information

between users. Since OSNs are internet-based networks, information could be, a form of multimedia content, a URL link and so on. Most studies in online information diffusion are focused in these information concepts.

3.1.1. Related Work

As previously mentioned, the topic of diffusion has been in the centre of sociology interest for decades. Even before the rise of Online Social Networks (OSNs), social ties and information flows have been studied, in traditional real-life social networks.

The notion that information flows, from mass media to opinion leaders and later to a wider population as final consumers, was firstly introduced during the middle of the 1940's [108]. In their introduction, Lazarsfeld et al., found that during a presidential election mass media had a direct influence on voting intentions, but informal and personal contact was more frequently mentioned as source of influence. Almost a decade later, Kaltz and Lazarsfeld revisited the subject by proposing their theory of the "Two step flow of communication" [109]. They proposed that the information from media is firstly received by opinion leaders who then pass their interpretation and actual information to individuals. In a similar conceptualisation, Granovetter and Mark noted that the analysis of social networks is suggested as a tool for linking micro and macro levels of sociology theory [4]. Small-scale interactions become large-scale patterns and feed back into small groups. They concluded that weak ties are seen as the individual's opportunity, towards integration into a community, while strong ties lead to overall fragmentation. Similar conclusions were verified in OSNs after nearly 40 years, while at the same time, the interest for social media analysis deepened [110], [111].

The term "Viral marketing" was introduced in 1997 by Juvertson and Draper [112]. In their work, the authors described their marketing strategy for a free email

service in which, by sending personal messages to individuals, boosted its user base. Their strategy proved to be very effective, gaining millions of users within a few months. During the following decade, not only e-mail but several internet-based services evolved (e.g. real-time interactive services, OSNs), becoming an integral part of marketing strategies worldwide.

Porter and Golan in [113] found that provocative content such as sexuality, humour, violence and nudity were crucial virality factors, compared to traditional TV advertising, where emotive content had always been the key factor. In addition to virality, information diffusion became an equally important research subject. Leskovec et al. [114] modelled the outbreak detection via node selection. They also performed a two-fold evaluation of their model, by using a water distribution network and a blog network. Although the described model was not verified on any OSN, it could be easily applied in such networks, mainly because OSNs have similar diffusion properties with the aforementioned networks.

Concerning word-of-mouth scenarios, Allsop et al. [115] noted that 59% of individuals frequently shared online content. The authors also tried to examine the link between emotion and virality, and concluded that content sharing was mainly adopted for entertainment purposes. Berger and Milkman [116] observed that positive content is more viral than negative. Moreover, they ended up with the conclusion that the more the content evokes emotions of activation (e.g. anger, awe, anxiety) the more viral it is, in contrast to deactivating emotions (e.g. softness).

The authors of [117] were focused on identifying the optimal network that best described information propagation in news media and blogs. News diffusion networks were found to have a core periphery structure, where a small set of core media sites diffused the information to the rest of the network, while blogs were mostly influenced

by mass media. In [118], authors analysed word of mouth through web means, mainly through email forwarding. Two dynamic patterns were observed, namely “Transmissibility” and “Fanout Cohesion”, while direct referrals proved the best form of diffusion, creating affinity paths.

Social influence modelling was the scope of [119]. The researchers used old propagation data to create such models and observed that viral marketing could leverage genuine influence, which -as they claimed- occurs only in real-world networks. Similarly, the authors in [120], investigated the social influence in meme social graph. They found out that most reports are made during the content creation period and shortly after it. Power users and post’s age were proved crucial virality factors.

In [121], the authors studied virality issues in Twitter. Due to its message-based nature, they observed that the direct mentions network was of higher value, while the information chains formed were topic-related and presented a short time span. Finally, they noted that content lifespan was very important to virality, also mentioned in [120]. In the same OSN, Sakaki et al. [122] studied users as social sensors, by monitoring information flow and dissemination during earthquake incidents. In their approach, they showed that earthquake detection can be equally effective compared to traditional monitoring techniques. In [56], authors analysed the entire Twitter graph in order to assess its topological characteristics. One of their findings was that each retweet would always reach 1000 users, regardless of origin or content. Twitter network had low reciprocity, with a certain level of homophily observed amongst its users.

In [123], Tucker measured the effectiveness of virality as a marketing tool. Unlike the work described in [112] -where direct emails messages were used- 400 videos and 24000 people were used to monitor information propagation. It was observed that for nearly every one million views, persuasiveness declined by 10%. However, this occurred

only after a negativity threshold, above of which that effect is observed. That threshold was 6 million views and was reached only by 24 specific videos. On the one hand, interaction affects were negative, while on the other hand, sharing engaging, provocative (as defined in [113]) or humorous content with visual appeal affected virality in a positive way.

In [124], authors created a map of the so-called “life cycle” of a blog. They found out that blogs have small cycles and can be split into groups, based on their content and role. Two of these groups, namely “elite” and “top political”, not only created political information, but they also drove and sustained the viral process. Likewise, a virality study in OSNs conducted by Guerini et al., showed that content is more important to virality than the influencer itself [125]. In this work, the authors used different metrics, such as buzz, appreciation, controversy and discussion raising.

Similarly, focused in the diffusion analysis of Twitter, Hansen et al. [126], they observed the effect of negativity in virality. More specifically, it was found that non-news positive sentiment and negative news, were more likely to be retweeted, thus supporting overall virality. They concluded with the following: “If you want to be cited: Sweet talk to your friends or serve bad news to the public”.

In other OSN-related studies, Bonchi [127] studied the maximization of influence, based on the most influential users of a particular OSN. The work described in [128] examined the passivity of Twitter users. Where it was found out that most of the users act as passive consumers, not forwarding news or content, thus a high popularity does not imply a high influence. They also argue that in order to maximize influence, one must overcome the passivity of the network. In [129], Mathioudakis et al. created an algorithm to describe influence propagation in terms of likelihood based on previous logs. They pointed out that sparsification is a fundamental data reduction operation.

Rajyalakshmi et al. [111], proposed a stochastic model for the diffusion of several topics. The authors discovered that strong ties played a significant role in virality, having homophily as a major contributor, also observed in [56]. Homophily stands as the tendency of individuals to bond with other individuals that have similar interests. Micro and macro scales of the network became relevant - similarly to [4] - and the authors noted that acts within groups have a global impact. Romero et al. [130], analysed hashtag diffusion in Twitter and discovered that initial adopters were fairly important. They also studied two concepts within Twitter. The first one was “stickiness”, as the probability of adoption based on exposure, and the second was “persistence”, defined as the continuing exposure to a hashtag. Their results validated the “complex contagion” principle of sociology.

Yang and Leskovec also studied Twitter [131], based on content exposure growth and fade over time. However, their focus was more on clustering and the discovery of distinct shapes of time. The proposed algorithm outperformed the simple K-Means algorithm in finding similarities between tweets and blog posts. The authors also distinguished 6 different temporal patterns and presented a model that can accurately predict attention pattern, using as little information as possible.

On a different, but very interesting basis, Guerini et al. [132] noted that dynamics of Social Networks exist in scientific literature. They used psycholinguistic analysis to determine abstract affect towards virality. Where, the linguistic style proved to be an important factor and papers with easier to read abstracts sections, were more frequently downloaded. Their study linked virality with linguistic style.

Going back to OSNs, Huang et al. [133], tried to effectively select a number of nodes in order to monitor information diffusion. They noted the importance of degree

centrality and out degree of a node. Additionally, they proposed a new node centrality method to identify the monitoring capability of a node, namely monitoring centre.

Guille and Hacid [134] studied the dynamics of the spreading process. Their research relied on the assumption that macroscopic level dynamics are explained by semantic, topological and temporal interactions in the microscopic level. The proposed model was based on machine learning techniques, as well as on the inference of time-dependent diffusion probabilities. Furthermore Bakshy et al. [110] addressed the problem of information diffusion in OSNs by performing a large-scale experiment of random exposement to information signals for 253 million users. The authors ended up with the statement -as seen in [4] and [56]- that strong ties are more influential but weak ties are responsible for propagating novel information.

Guille et al. in [134] also addressed the issue of information adoption in Twitter with respect to micro and macro level dynamics. They presented an adaptable graph-based prediction model that estimated and adapted its parameters using machine-learning techniques. Their results were similar with those described in [109].

In [135], researchers aimed to answer why and how some messages in OSNs become viral. They modelled information ageing and competing message streams. Their obtained results implied a threshold above which a message becomes viral. Moreover, they noted that competing streams further raise this threshold. Finally, Weng et al. [136] analysed longitudinal micro blogging data. The authors highlighted the importance of triadic closure and shortcuts based on traffic, regarding the evolution of the social network. However, triadic closure was relevant only in the early stages of a user’s lifetime, since after this stage user linking was generated by the dynamic flow of network information.

In the following sections, information diffusion in across multiple online networks, as well as the lifespan of the diffused information, will be examined. The focus will be in user-generated content. The provided results verify the perception of content and information connection in various OSNs. Giving the first proofs of multi-layered (multi-OSN) information flow.

3.2. Networks of Interest

The spark of interest towards studying the information diffusion on social networks, was a social news and entertainment site named Reddit, which is powered by user-generated content. Deviating from the widespread OSN structure where users are connected through various affiliations, users of Reddit are connected through their comments and common interests. Their user profiles are also different from every OSN. Instead of creating their personal profile based on their characteristics or interests, a user profile is formed with participation. Comments made, articles uploaded and every public activity is stored in their profile. In a way, every profile is the reflection of our interests in Reddit, which in turn could be considered as a projection of real life interests.

Content in Reddit is submitted in the form of a descriptive link and may contain: an image, a meme, a video, a question, an Ask Me Anything (AMA) session etc. Every interested user, can then vote and comment on that post. Posts that acquire a high enough vote ratio in a short period after their submission are moved to the front page[31]. In correspondence with votes, users that have created a post or commented one, gain or lose "karma", a vote-oriented unit of Reddit. Karma is used for user ranking and is the main factor of popularity within Reddit. Karma could be seen as the equivalent of friends and

[31] www.reddit.com, Last Retrieved: 13/01/2016

followers in Facebook and Twitter respectively. It is apparent that the Reddit community defines the content and determines its “success” or “failure”.

Interestingly, the “karma” distribution over Reddit users also follows a strong power law. One percent of active users have more than 20% of the total “karma”. Although the data used for these measurements was partially incomplete, it utilised a substantial portion of the active community[32]. A power law distribution that can possibly point to community bias towards popular users. User bias in OSNs is uncharted for these type of networks. Relevant studies have been only performed in offline social networks. The subject of bias in social networks, will be revisited in *Chapter 4*.

Concentrated in information diffusion and for the purposes of our research, we wanted to also use famous and heavily-visited OSNs, since inter-OSN diffusion was obvious in multiple instances. One of the most famous OSNs, Twitter, provided us with the ability to fully observe the impact of a front page Reddit post. Absolute and thorough collection of data was possible thanks to the front page content of Reddit being relatively fresh, and to the unobstructed access that Twitter API provides for recent tweets. We could also crawl older Twitter data but the API would restrict our results to a random group of at most 1500 entries.

In contrast, crawling Facebook and Google Plus didn’t prove so fruitful. Most of the content in these networks, is shared between friends in private groups. These posts are unreachable from an unsupervised crawler. Thus, we could only crawl public posts from both networks. As a result, only a fraction of diffusion was discovered on both networks. As if the public crawling limitation was not enough, Google Plus user engagement was

[32] http://www.reddit.com/r/dataisbeautiful/comments/27zyh6/karma_inequality_1_of_redditors_have_20_of_the/, Last Retrieved: 13/01/2016

substantially lower than Facebook and it was later decided not to include Google Plus in the study.

From the post and link analysis of Reddit, it was obvious that information is shared and spread between all kinds of online networks. Apart from the established OSNs, many Reddit posts link to multimedia hosting sites, as well as news agency and blog sites. One post in Reddit linking to these sites impacts their traffic substantially. But what is the impact of that exposure? How long after the creation of a post in Reddit does it take for the effects of its propagation to appear? How long does it take for that propagation to decay? Is this propagation dependant on the subject of each post?

These were some questions raised during our initial engagement with Reddit and its content. User-generated content and its topics combined with crawling and multiple OSN analysis could provide substantial data towards diffusion measuring. But the crucial element for a diffusion type of study is virality. The term viral is used to describe information that within a relatively small timeframe achieves worldwide or widespread diffusion. That type of content is substantial for the analysis of multi-OSN diffusion, for viral content spreads rapidly with considerably impact to linked networks. Thus, the focus of the next section is, the diffusion of viral user-generated content through multi-OSN analysis. The primary aim of this chapter is to present concrete proofs of the multi-layered flow of information in OSNs.

3.3. Lifespan and propagation of Information in Online Social Networks: A Case Study based on Reddit

As mentioned in the previous section, the initial platform for this information diffusion analysis is Reddit. To briefly recap, Reddit is an OSN hosting user-generated content. This content is ranked based on the resonance of the users which is derived by a voting system. The front page of Reddit comprises of posts that within a small timeframe have accumulated a high enough number of views[33]. Posts within Reddit can contain internal links or links that lead to external domains.

These external links are the cornerstone of multi-OSN diffusion. Not only do they spread within the initial OSN, Reddit in our case, but they also propagate to various other OSN and entertainment sites. This does not mean internal links are without any interest, only that they do not intensely propagate to other OSNs or domains in general. Most of the external links point to a hosting domain of multimedia content as Reddit is mainly used for entertainment. Apart from such domains, links could also lead to news or informative domains, such as www.cnn.com, www.bbc.com and www.wikipedia.org. But as not every link and its subsequent hosting domain proves useful for diffusion analysis.

3.3.1. METHODOLOGY

A post in Reddit includes a title and comments. The title section, apart from the obvious titling of the post, might contain an internal link that leads to a Reddit discussion thread or might link to an external site. Regardless of the title, each post also contains a comments section, where users can discuss the post. In addition, every post has to be submitted in a certain topic category, also known as "subreddit". Each subreddit is of different context, but not always strictly defined. For example, r/funny[34] is an

[33] http://amix.dk/blog/post/19588, Last Retrieved: 13/01/2016
[34] http://www.reddit.com/r/funny, Last Retrieved: 13/01/2016

entertainment subreddit, while r/AdviceAnimals[35] may be entertaining, emotional or informative. In our research, we used the external linking URL of the title to measure its diffusion and the corresponding subreddit to define its topic.

From a practical perspective, in order to accurately monitor and measure diffusion counters are needed. Whether these counters are IP-based or simple view counters is not of great importance. Of course an IP counter would also provide precise numbers, but would not account for the popularity of a given source within a network with one external IP, such as within universities and companies. External URL links pointing to domains that do not provide any kind of counter are not used in our monitoring process. In addition, any links that are not part of the post, such as links contained in the comments section of the original post, were not monitored.

Out of every linked domain that appeared in Reddit, a selection should be made based on monitoring ease. At the time of the research, summer of 2013, the most commonly appearing external domains were www.wikipedia.org, www.imgur.com, www.youtube.com and www.quickmeme.com. Out of those, only www.wikipedia.org did not provide some kind of counter. Furthermore, after creating our crawler based on www.imgur.com, www.youtube.com and www.quickmeme.com, the latter domain was banned from Reddit[36]. The ban was issued because Quickmeme domain used unfair practices to gain votes, which would eventually assist the post containing its domain link to reach the front page, another issue which will address in *Chapter 4*. But to our luck, the traffic and content volume from www.quickmeme.com shifted almost entirely to

[35] http://www.reddit.com/r/AdviceAnimals, Last Retrieved: 13/01/2016

[36] http://www.reddit.com/r/AdviceAnimals/comments/1gvnk4/quickmeme_is_banned_redditwide_more_inside/cao99l6, Last Retrieved: 13/01/2016

www.imgur.com. Thus, our crawler only needed a small modification in order to operate on two domains, namely on www.imgur.com and www.youtube.com.

As we mentioned, diffusion is directly connected to the virality of a post. In the interconnected environment of OSNs, virality almost acts as a warranty of diffusion. More views, generated by an increased interest, are bound to create traffic to most social networking sites. Due to the ranking algorithm of Reddit, virality is also connected with the age of a post. Thus, it was important to try and capture content that would eventually attain a high number of views in order to monitor from the initial submission, rise to popularity and its stableness.

Thereafter, we chose to crawl new and rising subreddits which both contain newly created posts. The New subreddit randomly picks new posts every 120 seconds, while the rising subreddit presents posts created in the last minutes with at least some upvotes The content of both subreddit contents are also refreshed every 120 seconds. For a two-month period, August and September of 2013, we crawled both of these subreddits. By selecting these two categories, we also fall in line with the API crawling requirements of Twitter, where we could reach an absolute retrieval for recent tweets, made within the last week.

During the 60 days of the crawling process, almost 1 million posts were obtained from Reddit. Each post belonged to a subreddit, which was used to denote its topic. Some of these topics were “pictures”, “gaming”, “funny”, “news”, “videos” and “music”. Unfortunately, not every topic was used in our dataset. As mentioned, only posts with external links to www.imgur.com and www.youtube.com were important to our diffusion analysis. Out of the 1 million crawled posts, only 102400 linked to either www.imgur.com or www.youtube.com.

3.3.1.1. Units of Interest

In order to evaluate the diffusion of information when dealing with multiple online networks and domains, a uniform unit of measure is needed. Page views, likes, tweet mentions and votes, were taken into consideration when defining that uniform measurement. One “Unit of Interest” (UoI) is strictly defined on actions that demonstrate interest of users throughout every domain. For each domain or OSN, UoI are portrayed differently.

So in Reddit, one UoI is equal to one vote. Although not every user votes on the submitted posts, voting is enough to project the community interest variations over time. Votes are probably more accurate in the monitoring of the initial interest, since they are factored in the ranking process of Reddit. In Facebook, one UoI is equal to one public post containing either the Reddit submission URL or the linked domain URL. Similarly, we also checked every last week’s tweet in Twitter, searching for the Reddit submission URL or the linked domain URL. One UoI in Twitter is defined as a mention of either URL. By “linked domain URL”, we mean the domain where the content used for the Reddit post is hosted. In our study, this could be either imgur.com or youtube.com.

For the domain UoI, the pairing is kind of obvious. In the ImgUr domain, one UoI is equal to one view of the hosted content, while in YouTube, one UoI is equal to one video view. Both view counters are easily parsed. However, during our result analysis we observed that www.imgur.com often encountered some down time, due to the large number of visitors, and would return 0 value when parsed. To surpass this, we used the average value of the previous and the next measurement in order to approximate the missing value.

3.3.1.2. The Virality Criterion

Not only fresh content was needed, it had to be viral as well. Viral content -as previously noted- is content that got a large number of UoI in a relative small timeframe. Virality, though, is not a standardized property by any means. On the contrary, it is a loosely structured concept of information diffusion and while many theories are proposed for its definition and subsequent calculation, none is universal adopted.

Thus, our own virality monitoring solution should be designed, based on our research needs. These needs were the processing speed and the sufficient post filtering. Since our crawler worked in a serialised fashion, the virality check needed to be executed almost instantly. A delay of 0.1 seconds for each of the 100000 posts in the check queue, would delay the monitoring process for almost 3 hours. As for the sufficient filtering and removal of non-viral posts, a strict rule would probably (and most certainly did) address the issue.

Our proposed virality criterion was a UoI variance check made in the first 4 checks of the Reddit UoI, after the creation of the post. If a given post would consistently –at least– double its Reddit UoI for each check, then it would be monitored throughout a 24-hour period. If not, that post was discarded as not viral. Although that did not mean the post would not eventually become viral or that it would not propagate to other networks. Our virality criterion is only a safe method to maximise both the virality probability of a post as well as its potential propagation to multiple domains and OSNs.

As expected, only a small number of posts passed this criterion. Out of the total 102400 linking to the domains of interest, www.imgur.com or www.youtube.com, only 682 posts doubled their Reddit UoI in the 4 initial checks. Which translates to 0.66%, of the total posts linking to the aforementioned domains.

3.3.2. THE MONITORING ALGORITHM

For each of the two subreddits, we independently executed a separate but identical monitoring algorithm. We started in the 1st of August 2013, at a predetermined time. Every submission in the respected subreddit was crawled. Creating a queue of *x* numbers of posts. Starting from the first post, we analysed the domain where the content of the post is hosted. If that content is hosted in www.imgur.com or www.youtube.com, then the post is added to the monitoring queue. If the post's content is hosted in another domain or within Reddit, then it is discarded. This process is performed every 120 seconds, which is the time both subreddits take to refresh their posts. This process was continuously executed for the entire two-month period.

In parallel to that process, each post from the monitoring queue was checked for every type of UoI. Again, from the first post to the last, we check for any UoI in Reddit, Facebook, Twitter and the hosting domain. So each post was tied to 4 separate UoI series and each check was separated by at least one hour from the next one. This time period was selected in order to overcome the request limitations of Twitter, since for every hour, a user is able to perform 350 API Twitter requests.

The first 4 checks are of great importance for the monitoring process. After the initial check and up to the fourth, we required the doubling of the Reddit UoI. Posts that didn't double their views were discarded from the monitoring queue. For example, a post that would have 3 Reddit UoI in the first check and 7, 18, 40 Reddit UoI for the next three checks would pass our virality check. While, a post with 3 initial Reddit UoI and 7, 15, 28 on the 3 subsequent checks would not pass the virality criterion.

Every post that passed the virality check was permanently added to the monitoring queue along with any new posts from the domains of interest. From that point onwards, that post was not bound to any UoI variance requirements. Interestingly, every

post of the 682 that passed the virality check acquired a large amount of viewership and climbed to the first page. This absolute discovery of viral posts points to a great selection through our virality criterion, which is also related with the ranking algorithms of Reddit.

After the 4th check and if a post has passed the virality criterion, the checks continue -as said- with at least an hour apart and for every check, all domains were crawled. Each post was monitored for at least 24 hours, with no less than 25 checks made for each submission. Almost all of the posts after the 14th check, which happened around the 18th and 24th hour, displayed a significant decay of interest in every domain and in every OSN.

3.3.3. RESULTS

The results of our crawling procedure were truly diverse. To capture the extent of that diversity, results were kept separated based on the subreddit of the post's submission, as well as the subreddit from where it was crawled and the domain it was linked to. The submission subreddit will be called topic, the crawled subreddit will be category, while the hosting domain will be referred as domain. Hence a post linking to www.imgur.com that was crawled in /r/new subreddit, but was submitted in /r/funny, would be classified as a post of the "ImgUr" domain from the "New" category in the "funny" topic.

3.3.3.1. Topics

By crawling "New" and "Rising" categories and "ImgUr" and "YouTube" domains, we came across a multitude of topics. Since Reddit is an OSN with a strong entertainment character the most prominent topic was "Funny". In "Funny" as well as in the "Gifs" and "Pics" categories, content -as expected- revolved around entertainment in the form of jokes, animated or still images, memes, sarcastic news and such. Similarly, the "AdviceAnimals" category is entertainment-focused with a small twist of actual life

advices. On the other hand, “WTF”, “Aww” and “EathPon” could be considered as amazement and awe evoking content. Posts in these categories usually present unexpected every day events, cute animals and beautiful external scenery, respectively. In another group, the “Movies”, “Music” and “Videos” categories contain multimedia content focused on popular movies and music bands. The “TIL” category is strictly informative, presenting snippets of information on commonly used websites that would be probably overlooked by users. Lastly, the “Gaming” category includes posts from (and around) the video games entertainment.

As with every OSN, users are interested in almost every topic corresponding to their real life interests. The detailed enumeration of the results with respect to the topic will further indicate the entertaining character of the Reddit community.

When directly comparing our two crawled categories, “Rising” and “New”, we found that the number of posts linking to the domains of interest (ImgUr or YouTube) were disproportionate. In particular, the “New” category contained way more posts linking to the domains of interest than “Rising”. A possible explanation would be the fact that current events and news gain attention relatively fast, thus they are featured in the “Rising” category. Most of these real life event posts could potentially be linked to several news agency domains, which were excluded from our diffusion analysis based on the lack of counters in theses news reporting domains.

Table 15*. Posts of Interest Domain Distribution*

	Posts of Interest	**ImgUr**	**Youtube**
Rising	44050	40966 (92.99%)	3084 (7.01%)
New	58350	51984 (89.08%)	6366 (10.92%)
Cumulative	102400	92950 (93.70%)	9450 (6.30%)

Nearly 45000 posts linking to “ImgUr” and “Youtube” were found in the “Rising” category. In the “New” category almost 60000 links were discovered containing links to “ImgUr” and “Youtube”. To be precise, the breakdown of results was as follows: out of the 44050 posts found in the “Rising” category 40966 linked to “ImgUr” and 3084 linked to “Youtube”, while in the “New” category” 51984 were found with content hosted in “ImgUr” and 6366 posts with hosted content in “YouTube”, out of 58350 posts.

In ***Table 15*** we can see the distribution of the discovered posts, with content hosted in any of our two chosen domains. Of the total 102400 posts, more than 93.7% were coupled to a photo or an animated image from www.imgur.com, and 6.3% of these total posts were created with a link to a video hosted in www.youtube.com. One of our main concerns was the large discrepancy between our two domains, but apparently OSN (and thus Reddit) users prefer the ease of conveying information through an image, rather than through a video. With exceptions applying on particularly interesting cases that account for a pretty small portion of the total “YouTube” content.

With respect to virality, as defined by our criterion, our results showed an increased number of viral posts for content featured in the “New” category. Concerning Reddit submissions with content hosted in ImgUr from the “New” category, number of posts that passed our virality criterion was double compared to content in “ImgUr” from the “Rising” category. This lower virality seen in posts of “Rising” category was unexpected, because the UoI count in the first check was considerably higher than posts from the “New” category.

To be precise, in the “Rising” category out of the 40966 posts linking to “ImgUr”, only 200 succeeded in our virality criterion, a mere 0.48%. While for the 3084 posts including “YouTube” content in the same category only 14 went through our criterion, a bit more than 0.45%. On the other hand, in the “New” category and for posts with links to

“ImgUr”, which were 51984, the respective virality criterion success percentage was around 0.82% with 431 posts. Lastly, in the “New” category of “YouTube” linked submissions, of the total 6366 discovered only 37 passed our virality check, slightly higher than 0.58% success percentage.

It is worth noting that the proposed virality criterion is not an absolute nor a universal virality validator. It merely serves as a virality indicator which ensures that the post in check was viewed by a significant number of users in the small timeframe of at least 4 hours, maintaining the relative interest of the users. In total, 682 posts successfully passed the virality criterion. Of them, 631 were using content hosted in “ImgUr” while 51 were using some form of multimedia content hosted in “YouTube”. A further separation based on the category, revealed a total of 214 posts in “Rising” and 468 submitted in “New”.

Table 16*. Posts sorted by Topic, Category and Domain*

Topic	**Advice Animals**	**Aww**	**Eathpon**	**Funny**	**Gaming**	**Gifs**
ImgUr Rising	27	18	2	84	25	4
YouTube Rising	-	-	-	-	-	-
ImgUr New	73	38	5	158	69	8
YouTube New	-	-	-	-	-	-
Topic	**Movies**	**Music**	**Pics**	**TIL**	**Videos**	**WTF**
ImgUr Rising	3	-	29	-	-	8
YouTube Rising	2	1	-	1	10	-
ImgUr New	5	-	56	-	-	19
YouTube New	2	3	-	1	29	2

Although the topic separation was uneven between categories, it was spread throughout every available topic. The accurate values sorted by their hosting domain, the submitted category and the topic of the submission are depicted in **Table 16**. As mentioned in the previous section, entertainment posts were the most pervasive ones, with 439 posts. Gaming posts were second with 94 posts, while posts that evoked emotions were third with 92 posts. Posts based on multimedia content were fourth, with 55 entries. Lastly, there were only 2 informative posts.

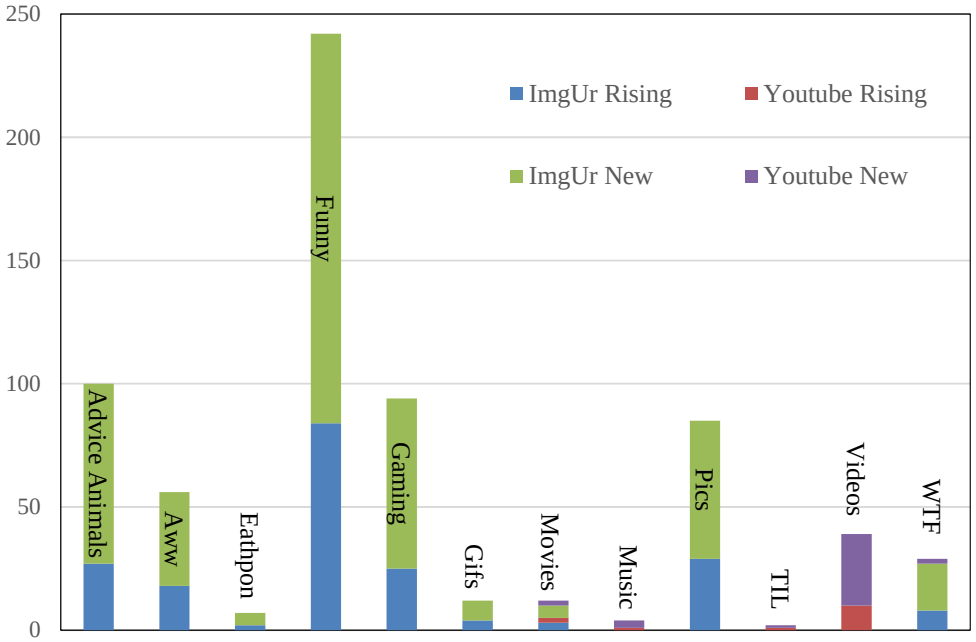

***Figure 19**. Posts per Topic*

To further visualise the disparity of the viral posts based on the proposed virality check, ***Figure 19*** is presented in a bar chart format. In it, we can distinguish the dominance of the “Funny” and “AdviceAnimals” categories as well as the lack of multimedia content and the absence of informative posts. The results of [9] and [10], that showed positive and entertaining content as the most frequent shared topic among online users, also apply in Reddit. More than 64% of the posts that succeeded in doubling their

Reddit UoI within the first four checks, were originating from entertaining categories. The gaming posts volume, which was the third highest per topic at almost 14%, provides a correlation of the user interest within Reddit and the general interest in video games. This is evident in the 6 million members of “Gaming” subreddit. Emotive content, as described in [7] and defined by the separation of topics, follows with a relatively small number of posts.

3.3.3.2. OSN Diffusion

Out of the 682 posts that went viral within Reddit, based on the proposed virality criterion, 550 were mentioned in at least one OSN, somewhat more than 80%. A mention -within this study- in any OSN, means providing either a URL of the Reddit submission or a URL the submission’s content in its hosting domain. For the “Rising” category, 114 post mentions were made in Twitter and 2 in Facebook, while 59 posts were mentioned in both. Similarly, in the “New” category, 257 mentions were discovered in Twitter and 5 in Facebook, whereas 113 post mentions were made in both OSNs.

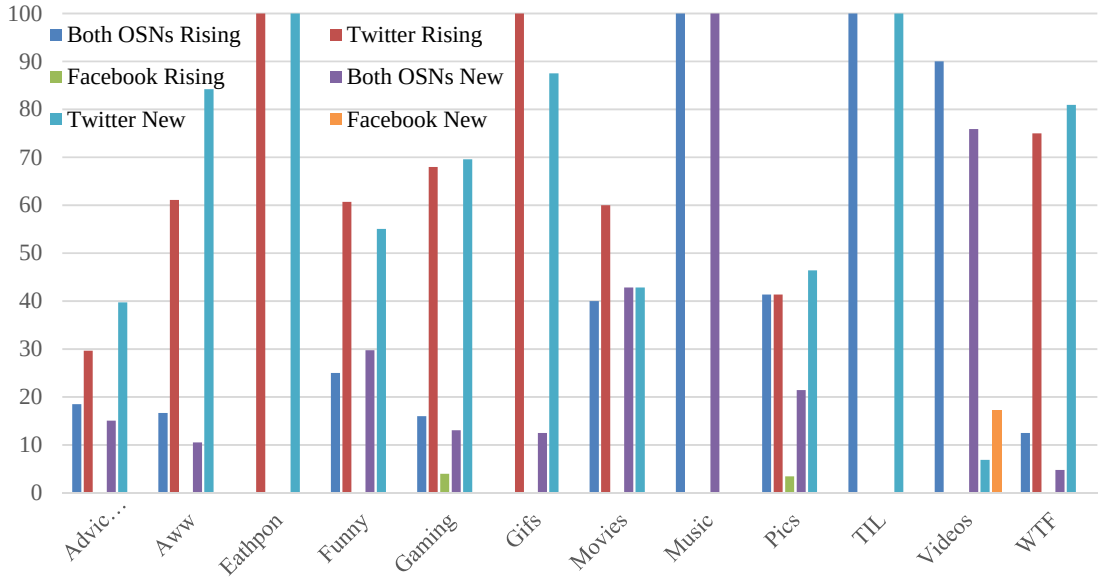

Figure 20. *Percent Diffusion of posts*

An analysis of the diffusion findings per topic, further solidifies the virality success results. The detailed counting of post mentions per topic is presented in ***Table 17***. The "Funny" topic is the most frequently mentioned in OSNs, closely followed by "Gaming" related posts and several other entertainment topics, such as "AdviceAnimals", "Aww" and "Pics". This diffusion spread is visualized in **Error! Reference source not found.** showed that almost 25% of posts are mentioned in both OSNs, while more than 54% are mentioned solely in Twitter, with only 1% mentioned only in Facebook.

Table 17. *Post mentions per Topic*

Topic	**Rising Category**			**New Category**		
	Both OSNs	**Twitter**	**Facebook**	**Both OSNs**	**Twitter**	**Facebook**
AdviceAnimals	5	8	-	11	29	-
Aww	3	11	-	4	32	-
Eathpon	-	2	-	-	5	-
Funny	21	51	-	47	87	-
Gaming	4	17	1	9	48	-
Gifs	-	4	-	1	7	-
Movies	2	3	-	3	3	-
Music	1	-	-	3	-	-
Pics	12	12	1	12	26	-
TIL	1	-	-	-	1	-
Videos	9	-	-	22	2	5
WTF	1	6	-	1	17	-
Total	**59**	**114**	**2**	**113**	**257**	**5**

The percentage of the diffused posts over the viral posts, where users' sharing habits are clearly seen, is presented in ***Figure 20***. Almost every topic displays a diffusion

percentage of 40% or above, in at least one OSN. A small minority presents this diffusion behaviour throughout both categories. Whilst, Facebook-only mentions are almost non-existent in every topic of the “New” or “Rising” categories. Multimedia content is shared to OSNs in percentages close to 100%, but the number of posts with multimedia content is not sufficient to generalize that rule.

The “New” category diffusion, is shown in ***Figure 21***. Twitter was the most commonly used OSN, for sharing content of Reddit. The most shared topics (percent wise) in Twitter were “TIL” and “Eathpon”, followed by more than 70% of posts with topics such as “Aww”, “Gaming”, “Gifs” and “WTF”. Posts from topics: “Funny”, “Movies” and “Pics” were round the 50% sharing mark. On the contrary, the least shared topic in Twitter was “Videos”, but it was the only topic shared solely in Facebook. As for both OSNs diffusion, every topic sharing percentage was way lower than 40% except from “Music”, “Videos” and “Movies”, once again dictating a preference of users towards multimedia content in both Twitter and Facebook. “AdviceAnimals” has a really low total sharing percentage, with 15.07% and 39.73% of its posts shared in both OSNs and only Twitter respectively.

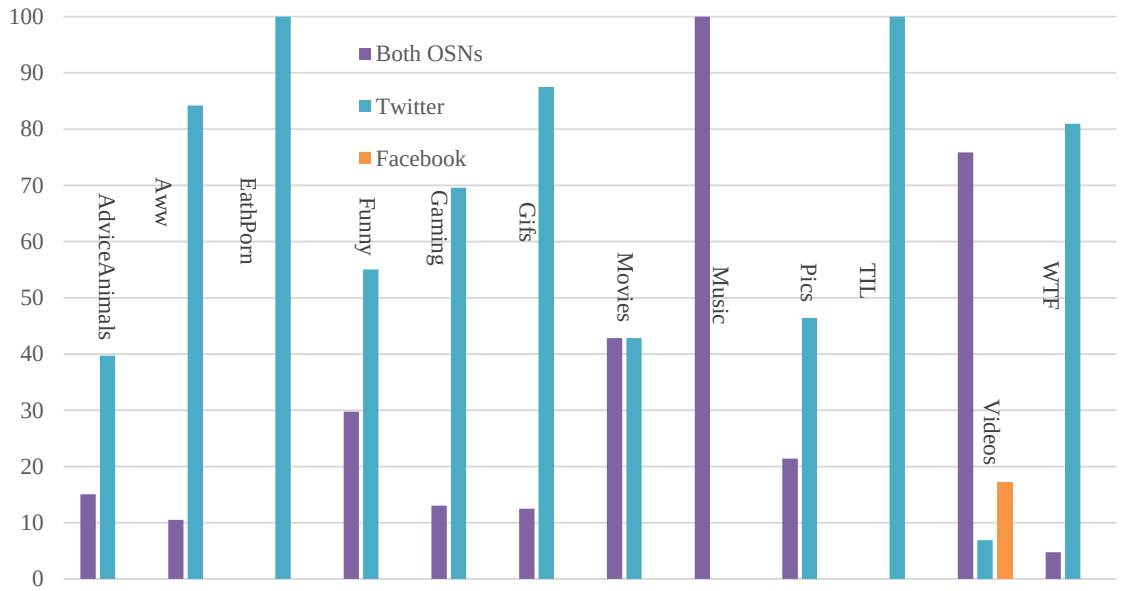

Figure 21. *"New" category diffusion percentages*

A similar dispersion appeared in the "Rising" Category, as seen in ***Figure 22***. The entertainment content was mostly shared in Twitter, while multimedia content was solely shared in both OSNs. Facebook-only sharing, once again, was very low and applicable only to the "Pics" and "Gaming" categories. "Eathpon" and "Gifs" were mostly shared in Twitter. Similarly, "Music" and "TIL" were 100% shared in both OSNs and Twitter respectively closely followed by "videos" with a 90% sharing in both OSNs. In this category, only "WTF" posts were shared over 70% in Twitter, with "Gaming", "Aww", "Funny" and "Movies" topics sharing percentages, closer to 60%. "Pics" topic was it is shared in both OSNs as frequently as only in Twitter, similarly to "Movies" topic in the "New" category. Finally, "AdviceAnimals" was once again the least shared topic of the "Rising" category, with 18.52% of its posts shared in both OSNs and 29.63% shared in Twitter.

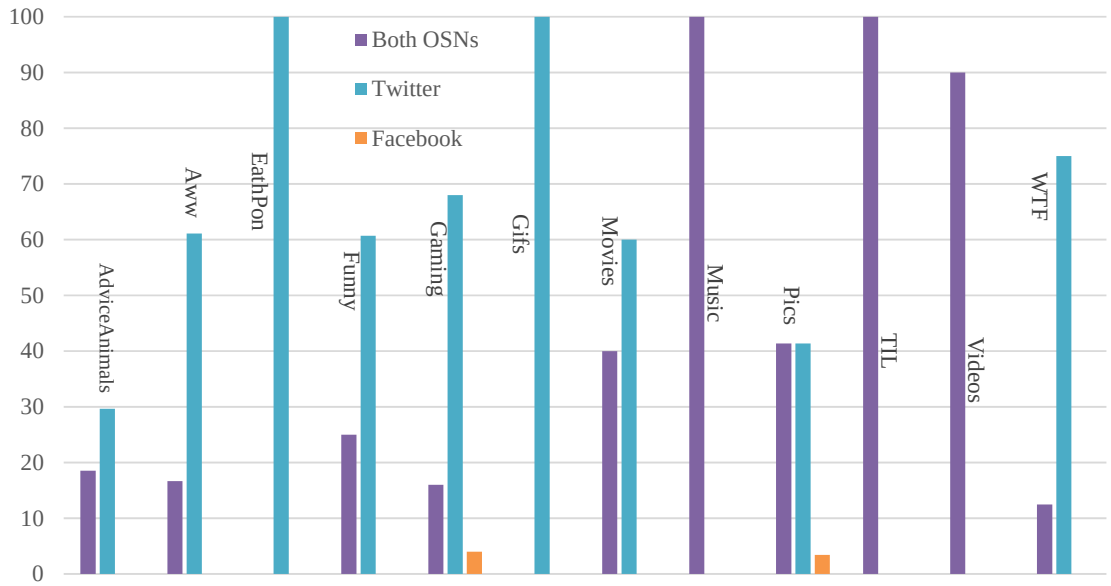

***Figure 22**. "Rising" category diffusion percentages*

The aforementioned results should be considered in correlation with the quantity and the quality of the crawled data. In ***Figure 21*** and ***Figure* 22** we present the percentages of OSN diffusion for posts that passed the proposed virality check. However, not every category had a sufficient number of posts to ensure the reliability of our results. The categories with a number of posts lower than 10 were "Eathpon", "Gifs", "Movies", "Music" and "TIL". On top of that, Facebook and its API crawl limits which forbid access to private posts, cannot accurately capture the interest pulse of users. Thus, we encourage readers to address the aforementioned categories and Facebook mentions in the following sections, as merely indications of diffusion and by no means not definitive conclusions.

Table 18. "New" category diffusion percentages

Topic	**New Category**		
	Both	**Twitter**	**Facebook**
AdviceAnimals	15.07	39.73	-

Aww	10.53	**84.21**	-
Eathpon	-	100	-
Funny	29.75	55.06	-
Gaming	13.04	69.57	-
Gifs	12.5	87.5	-
Movies	42.86	42.86	-
Music	100	-	-
Pics	21.43	46.43	-
TIL	-	100	-
Videos	**75.86**	6.9	17.24
WTF	4.76	80.95	-

In order to be more precise with the diffusion percentage values, ***Table 18*** and **Table 19** are included. The greyed out lines denote the topic in which the number of posts was exceedingly low. While the max percent post diffusion on each OSN column, excluding the greyed out lines, can be seen in bold. The absence of diffusion is indicated with the symbol "-", in both tables.

Table 19*. "Rising" category diffusion percentages*

Topic	**Rising Category**		
	Both	**Twitter**	**Facebook**
AdviceAnimals	18.52	29.63	-
Aww	16.67	61.11	-

Eathpon	-	100	-
Funny	25	60.71	-
Gaming	16	68	4
Gifs	-	100	-
Movies	40	60	-
Music	100	-	-
Pics	41.38	41.38	3.45
TIL	100	-	-
Videos	**90**	-	-
WTF	12.5	**75**	-

3.3.4. DIFFUSION OVER TIME

As mentioned in *Section 3.3.2*, our algorithm was designed in order to monitor diffusion over time of a certain Reddit post. During the monitoring period, UoI checks were performed in –at least– an hourly basis for Reddit, Twitter, Facebook and the hosting domain ("ImgUr" or "YouTube"). We monitored the propagation of a post within our designated online environments in frequent time intervals.

Slow reaction times and reduced viewership outside of the initial submitted OSN were some of our expectations, mainly due to the "tightly knit communities" effect of social networks as observed [137], [138] and [139]. Such patterns were commonly observed in various topics along with additional sparse and contradicting patterns in other topics. For each topic and hosting domain, the respective diffusion over time will be presented.

In the following charts, the mean UoI percent variances per topic for each of the OSNs and the hosting domain are used. Posts that did not propagate to Facebook or Twitter are not used at all, even though they passed the virality criterion. This means that at least some level of UoI variance was required for the posts to be included in the calculations. The UoI variance is based on the UoI checks we conduct. For each check after the first, we measure the growth in UoI in the current state compared to the previous

one. For example, if during the first check we discover 10 Reddit UoI and we find 30 Reddit UoI in the second check, the Reddit UoI percent variance is 200%. It is calculated by:

$$Percent\ Variance = \frac{Current\ UoI - Previous\ UoI}{Previous\ UoI} \times 100\ ,$$

where current UoI and previous UoI refer to the UoI checks, and the UoI could be within Reddit, Facebook, Twitter and the hosting domain.

These UoI checks are performed in short time intervals for each post, which as mentioned were at least one hour apart, and every check monitors the UoI in the hosting domain, Reddit, Twitter and Facebook. Each post has a maximum monitoring period of seven days, after passing the initial virality check. Once more, this is due to Twitter API access limits, which do not allow accurate crawling for tweets more than seven days old.

It was observed, for all post afters the 14th check, user interest decayed on every OSN and domains. The 14th check happened somewhere in between the 14th and the 24th hour of submission. The mean number of checks for each post were 65 and was heavily influenced, by the serialised nature of our algorithm and the low number of checks of the newly discovered content.

Bound to the virality check and the monitoring process, which needed Reddit UoI to be doubled in every check, all posts presented a high Reddit UoI percent variance for the first four checks. After the fifth check, few posts maintain a significant level of interest throughout every OSN and domain. Facebook API limitations are also affecting the results and the visualization of the UoI percent variance. Based on that percent variance over time, most topics can be categorised into distinct patterns. Patterns that are identifiable over the first 14 checks, when interest is vibrant. According to our results, the

percent variance after the 14th check was less than 1%, a practically insignificant value on samples that have accumulated tens of thousands UoI.

The pattern charts are separated based on the topic and the category. Multiple topics could follow the same pattern. This patterns will be presented in detail. In the following charts UoI percent variances over time are presented in an x-y fashion. The X axis corresponds to the number of check, starting from the initially crawled point up to the 14th check. The Y axis indicates the UoI percent variance as defined by the previous equation. This chart layout allows for a better identification of UoI spikes and decline contrary to other possible visualization methods (e.g. the use of logarithmic y axis).

3.3.5. DIFFUSION PATTERNS

Five distinct propagation patterns were observed. The topic classification, based on the observed diffusion pattern, is presented in ***Table 20***. Cells with grey shading match topics with less than 10 posts, whereas topics that had both categories following the same diffusion pattern are written without any category indicator. For example, both the "AdviceAnimals" topics from the "New" and the "Rising" categories, followed the Pattern 1 propagation. Thus, we only used the topic title in ***Table 20***. Contrary to "WTF", where posts from the "Rising" category followed Pattern 1, posts from the "New" category followed Pattern 4. Both occurrences of "WTF" are noted with their respective category, which also applies to "Movies" Topic.

In ***Table 20***, the first three patterns correspond to content hosted in "ImgUr", while patterns 4 and 5 were encountered in content from "YouTube". In addition, the only content topic that had submission in both "YouTube" and "ImgUr" domains, was "Movies". Which also had a unique diversity in its pattern breakdown. "Movies" content with picture or animated pictures hosted in "ImgUr" had a completely different diffusion from "Movies" video content of "Youtube". Although "Movies" articles from "ImgUr"

had a similar diffusion pattern for both categories (“New” and “Rising”), the multimedia “Movies” content from “YouTube” had a distinctly different propagation pattern for each category.

***Table 20**. Propagation Patterns*

Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5
AdviceAnimals	Movies	Gaming	TIL	Videos
Aww				
Eathpon			WTF New	Music
Funny				
Gifs			Movies New	Movies Rising
Pics				
WTF Rising				

3.3.5.1. Pattern 1

Upon examining original content created purposefully for a specific online network, it is expected to encounter high viewership, initially within the originating online network and potentially followed by any other neighbouring networks. That was the case with the observed Pattern 1. The mean UoI variance of the 13 different topics and categories for the first 14 checks, is presented in ***Figure 23***. The 7 different topics that were used to calculate the mean UoI variance can be seen in ***Table 20***, in the first column under the “Pattern 1” cell.

During the first check, all three OSNs and the hosting Domain had a given number of UoI. From the second check up to the fourteenth, we calculated the UoI variance of the checked post, using the equation found in *Section 3.3.4*. Thus in the first check, when the initial UoI were parsed, the variance was zero in the four Domains. In

the second check, we can see the Domain UoI (purple line) grew by 1200% followed by a 500% growth of UoI in Reddit. While in the third check, the Reddit and Domain UoI slowly fall, Twitter UoI rose by almost 100%. In the following fourth check, Twitter interest growth slowly faded but Facebook UoI started to increase, until peaking at the fifth check. The interest growth in the OSNs and the hosting Domain persisted until the fourteenth check. At this point, the UoI variance was less than 5% for the Domain and in every OSN. To further improve the presentation of the declining interest, which corresponds to a decline in UoI growth, ***Figure 24*** was included.

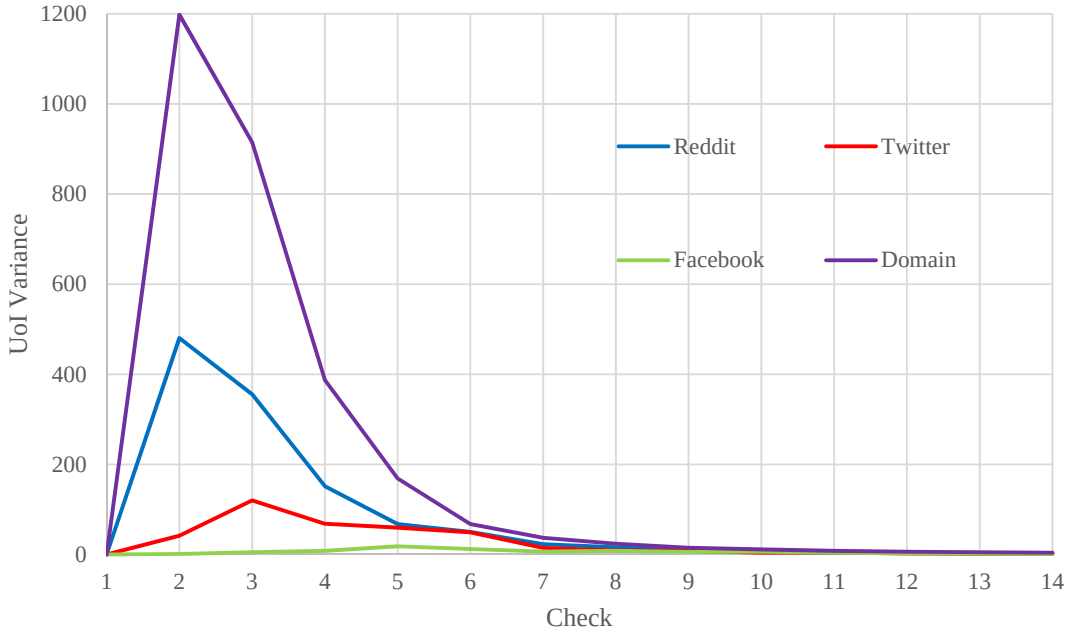

Figure 23*. UoI variance per check*

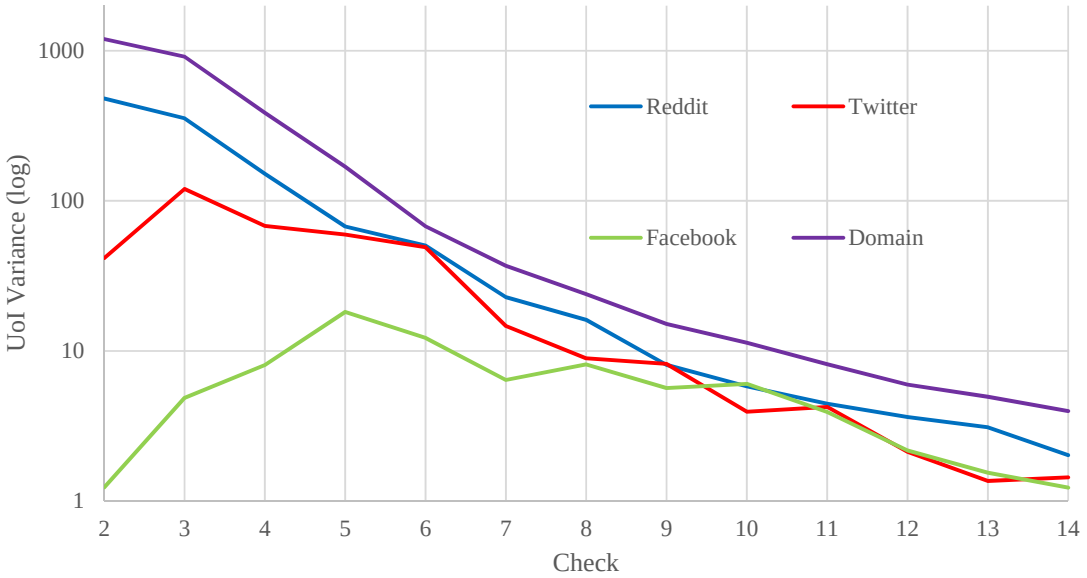

***Figure 24**. UoI variance per check, log scale*

The initial UoI count on both Reddit and ImgUr was almost identical, contrary to their UoI variance. This phenomenon could be interpreted by the usage and by the external viewership of each network. First and foremost, the hosting Domain "ImgUr" is one of the most visited Domains in the world, ranked 45 as of November 2015[37]. It is linked and viewed by a plethora of sites and users. Taking into account that most of these sites are also viewed by a broader audience and that a significant amount of their users would not visit the original source, the reason for that UoI variance difference between Reddit and the hosting Domain becomes apparent. Twitter, which is more of a content sharing -than content creating- OSN, also has a significantly lower UoI initial count and variance. Finally, Facebook is the most mainstream OSN of the three we tested, but had the lowest UoI initial count and variance. An observation bound to the API restriction that only allowed public posts to be crawled.

[37] http://www.alexa.com/siteinfo/imgur.com, Last Retrieved: 13/01/2016

We should keep in mind that when the UoI variance tends to zero, this is only an indication of a halt of interest growth rather than of a halt of pure user viewership. When a post has substantial hundreds of thousands UoI, even a 1% growth is an accountable number of UoI. UoI variance is measuring public's interest in a given post not the exact UoI count. These remarks are applied in every introduced pattern.

3.3.5.2. Pattern 2

The second Pattern was observed in "Movies" post from both the "New" and the "Rising" categories. Usually, the still-image content posted about movies is a screen capture, a photograph or a poster, of an upcoming film. That type of content, such as information and previews of upcoming titles, are time-sensitive. The mean UoI variance of posts from "Movies" topic-from both categories, is presented in ***Figure 25***. In ***Figure 26*** we present the corresponding log values of the UoI variance. At a first glance, the diminishing of interest after 6th check is obvious.

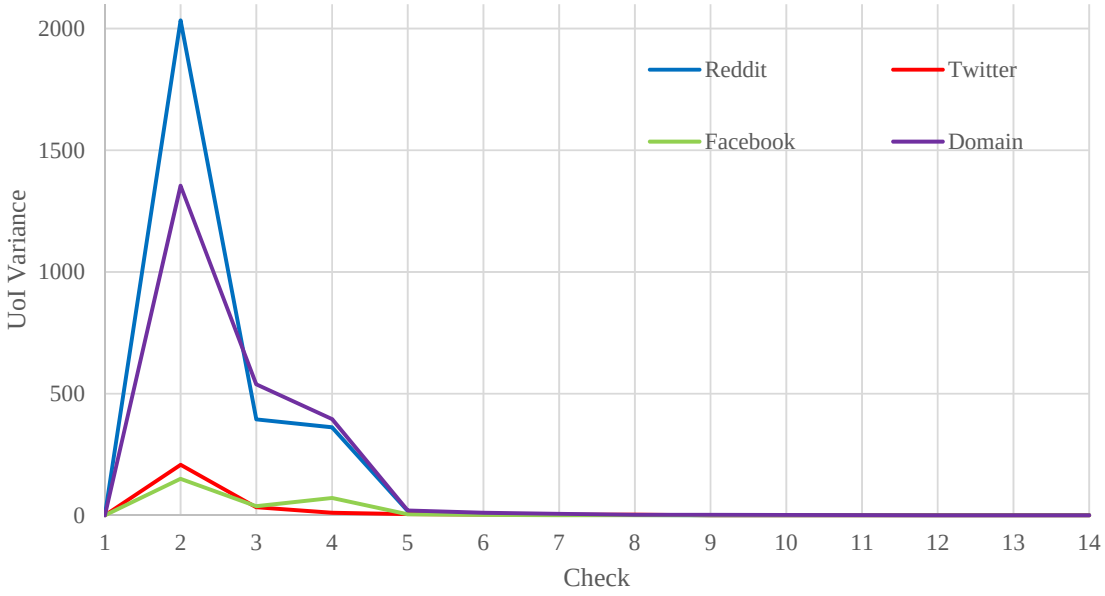

Figure 25*. UoI variance per check*

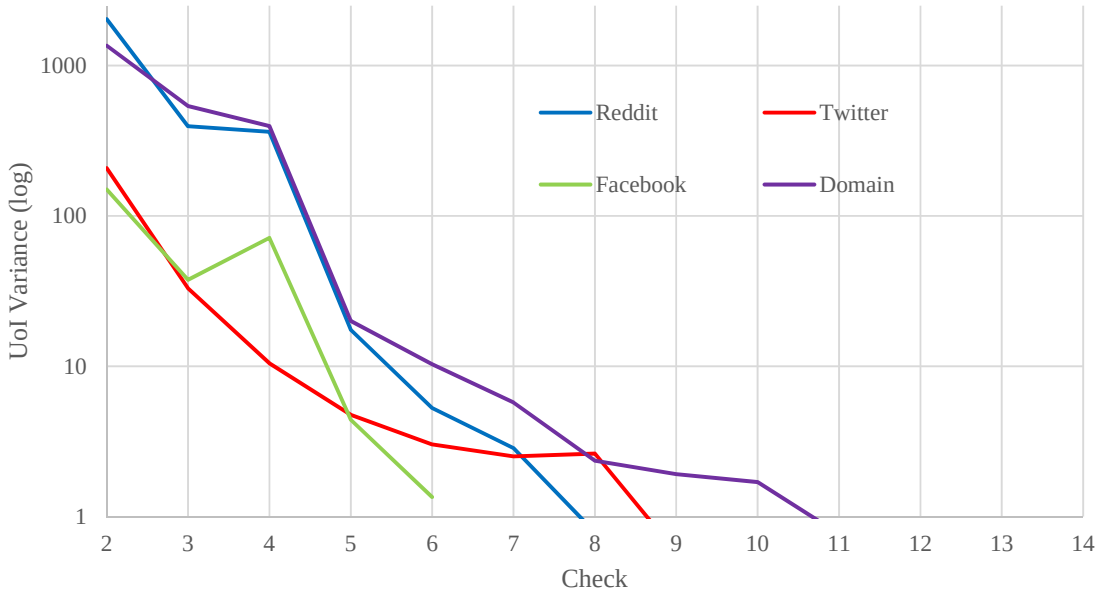

***Figure 26**. UoI variance per check, log scale*

The focus on this Pattern is the concurrent UoI interest spiking, in all the three OSNs and the hosting Domain "ImgUr" as well as the sudden loss of interest. In the second check, all the networks present a great UoI variance. In addition, contrary to Pattern 1 where the Domain UoI had a significant larger growth, the Reddit's UoI variance is way higher than the Domain's. Twitter and Facebook also present their highest UoI per cent growth during the second check. However, all the UoI variances drop significantly after that point, with only Facebook and Twitter presenting a small UoI increase in the third and eighth checks respectively. These events, the diminishing of interest and the increases in the UoI variance, can be clearly observed in ***Figure 26***. The second point of interest in Pattern 2 is the sharp decrease in the UoI variance. For Twitter, the complete lack of UoI variance is taking place in the seventh check, one check after Reddit and two checks before the Domain UoI variances come to a halt. Even though the Domain UoI growth lasted longer, it was equal to zero by eleventh check.

One of the possible explanations for the short lifetime of such posts is time relevancy, i.e. for how long is a new “movies related” image relevant. For example, if a screen capture of a movie is posted, it becomes obsolete when the next type of content is introduced, which may be an interview or a trailer video from the same movie. This effect is even more present, in cases when the content posted after a screen capture or a photograph is in multimedia form. This was the behaviour of the monitored posts from “Movies” topic. As for the low Domain to Reddit UoI ratio, the most probable reason is the contents’ origins. Most of the Reddit posts with Hollywood movies content –hosted in “ImgUr”- do not constitute the original content source, but this type of content is usually provided by other sources which are equally spread through every OSN, since movies are the most common and widespread form of entertainment.

3.3.5.3. Pattern 3

“Gaming”, which is another very popular entertainment subreddit among Reddit users, also revealed a particular Pattern in both categories. Unlike Pattern 2, where the subject was very time sensitive and widespread, gaming does not constitute a typical form of entertainment. However, “Gaming” UoI variation was more persistent than “Movies”. In ***Figure 27*** and ***Figure* 28**, we present the UoI variance and the log UoI variance respectively.

As seen in Pattern 1, Domain and Reddit UoI spike simultaneously in the second check, with the Domain UoI having a significantly higher variance, but the main point of difference is the Domain UoI variance. Which rises higher than the Reddit UoI variance in the sixth and the eleventh checks. Moreover, the Facebook UoI variance has a different behaviour from Pattern 1. Not only does it spike two checks before Twitter, it also presents an interesting overtake of the Twitter UoI variance, in checks eight and eleven.

Both the Twitter and Facebook UoI variances had a small increase during the fourteenth check. In this pattern, OSN interest is higher and of broader duration than the one met in Pattern 1.

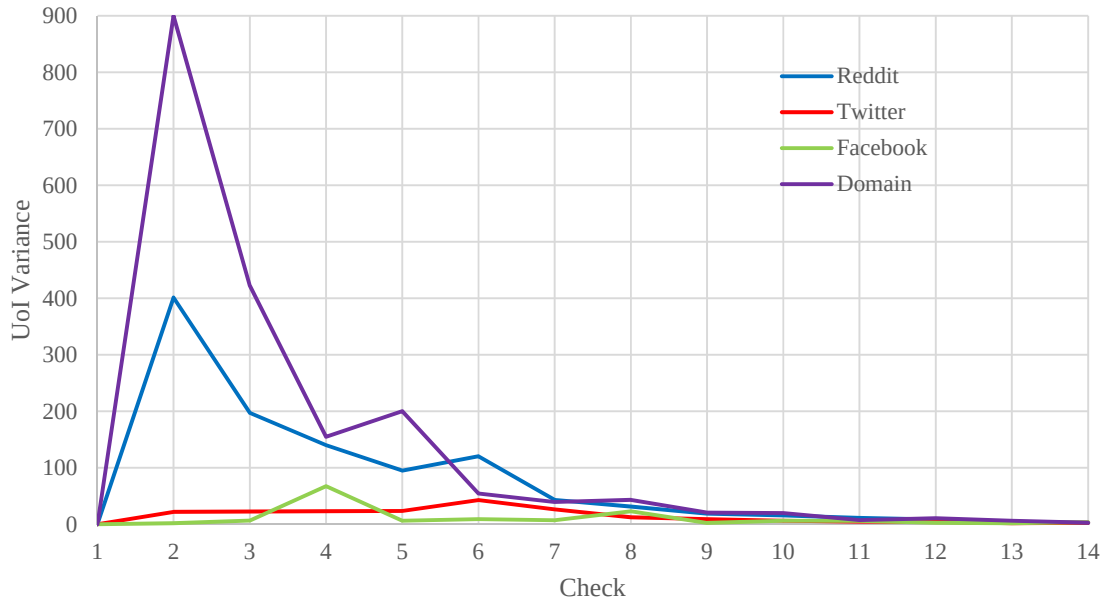

***Figure 27**. UoI variance per check*

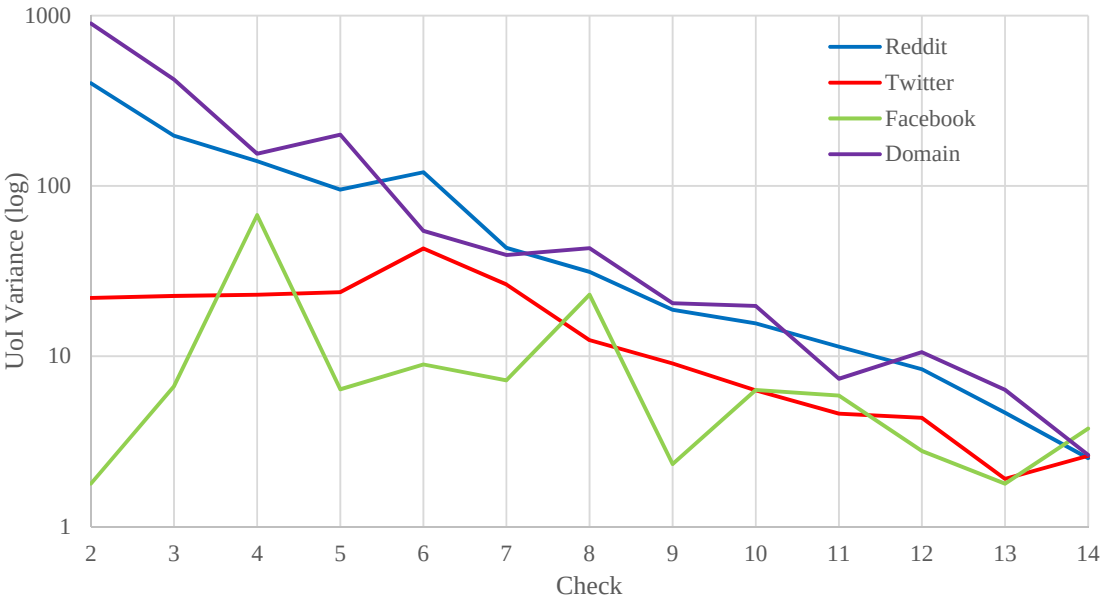

***Figure 28**. UoI variance per check, log scale*

Our analysis of the posts from "Gaming" topic revealed a particularly persistent interest of users within every OSN. One probable cause for the reach of "Gaming" posts could be the popularity of the corresponding subreddit, which is the seventh larger in Reddit[38]. Furthermore, it seems that users with these particular entertainment interests are highly active and even more prone to sharing and discussing relevant content. "Gaming" topic was also found at the fifth place of the most commented topics in Reddit[39].

3.3.5.4. Pattern 4

Posts that were hosted in "YouTube" they had completely different propagation patterns. Although n most of the topics had very few posts that passed the virality criterion, the UoI variance was fairly consistent amongst these few posts. The UoI variance of Pattern 4 is shown in ***Figure 29***, along with the UoI variance in logarithmic scale in ***Figure 30***.

[38] http://redditlist.com/, Last Retrieved: 13/01/2016

[39] http://www.reddit.com/r/TheoryOfReddit/comments/1qyvlp/update_is_there_a_method_to_figure_out_which/, Last Retrieved: 13/01/2016

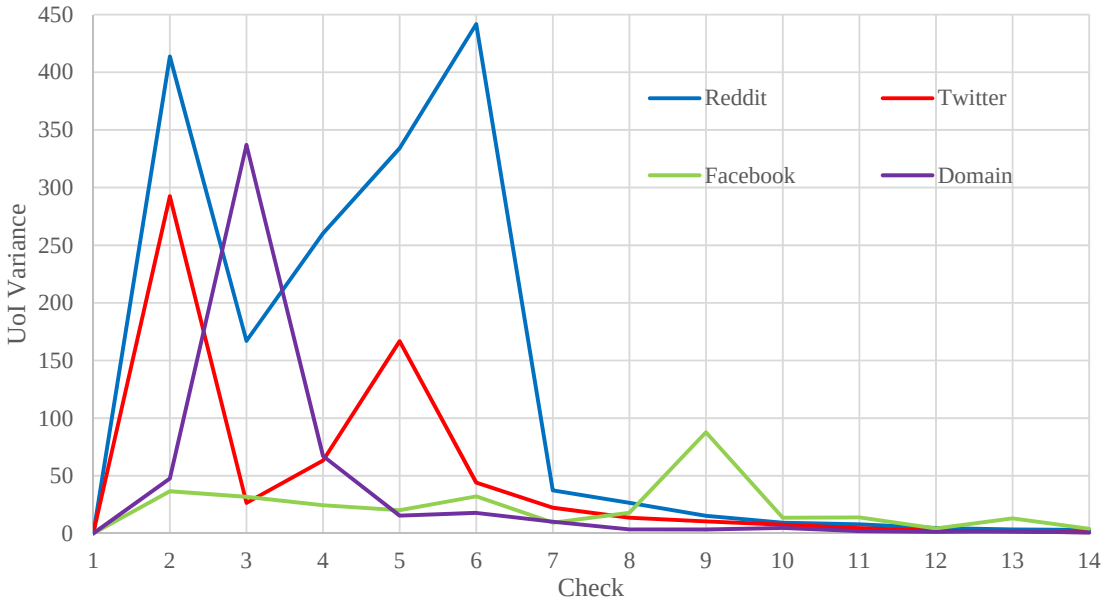

***Figure 29**. UoI variance per check*

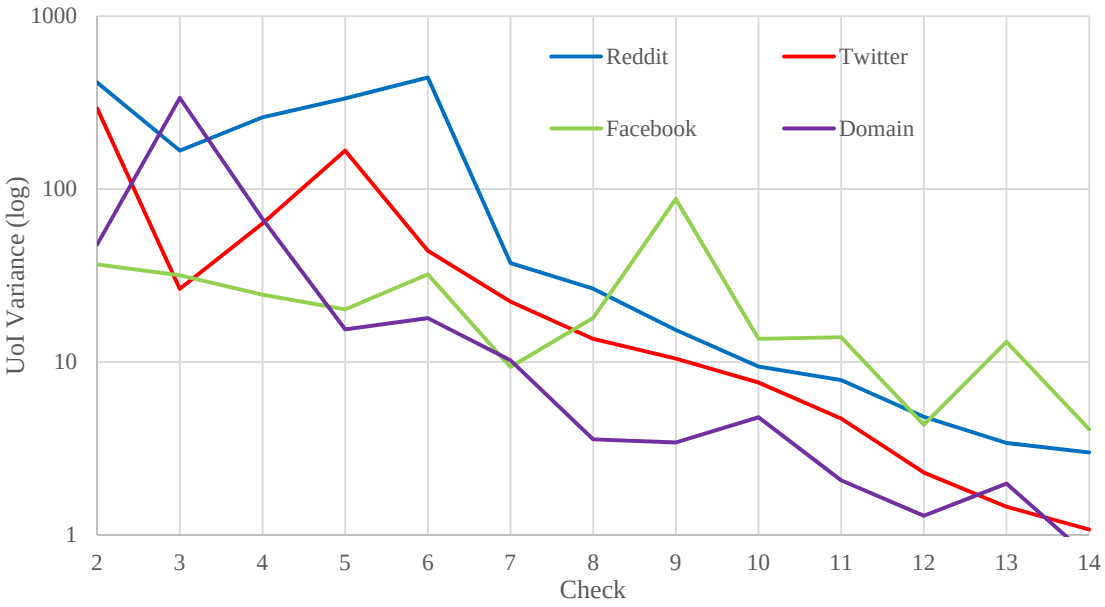

***Figure 30**. UoI variance per check, log scale*

A very different Pattern, compared to the previous three, appeared for these "YouTube" related posts. At the second UoI check, the Reddit and the Twitter UoI grow substantially, while the Domain and the Facebook UoI only grow by a small margin. The

Domain's UoI increase considerably only during the third check. From the fifth check and onwards, Domain's UoI variance is the lowest out of every monitored network. At the same (fifth) check interest rises for the second time within the Reddit, the Facebook and the Twitter. During the sixth check, by the largest UoI variance growth of Reddit. At the ninth check, both Reddit and Twitter users slowly lose interest, contrary to Facebook, where UoI are multiplying and peaking. Once again, opposed to the first three Patterns observed, the Facebook UoI variance is above every other from the ninth to the fourteenth check, only briefly surpassed by the Domain's UoI in the twelfth check. Overall, the UoI variance decreases with a similar pace to that of the first Pattern, but with greater OSN UoI variance rather than the Domain one. Interest spikes are easily identified in ***Figure 29***, whereas the variances per network could readily be compared in ***Figure 30***.

This UoI distribution and variance, is typical for videos that were not created specifically for Reddit. These videos are usually created before any respective submission in Reddit, but they enjoyed a late popularity after the posts appeared in Reddit or Twitter. This explains the relatively low Domain UoI variance as well as its steep increase and decrease. The Reddit and the Twitter interests lead to Domain UoI, which in turn refuels the interest in the initial OSNs. The Facebook interest increased a little late but it sustained a high variation until the end. We should though, keep in mind that both ***Figure 29*** and ***Figure* 30** present the mean UoI variances of less than 10 posts.

3.3.5.5. Pattern 5

The last Pattern is also based on "YouTube" linking posts. Pattern 5 is vastly different from the three "ImgUr" Patterns but it shares some characteristics with the fourth one. The number of posts that passed the virality criterion (and were observed with a similar UoI variance) is almost ten times the number of the posts met in the 4th Pattern.

Figure 31 and ***Figure* 32** present the mean UoI variance of these posts over check iteration in non-log and log Y-axis respectively.

It is obvious that external viewership of Domain UoI for “ImgUr” and “YouTube” content affects the Pattern 5 variance distribution. With the term “external” we describe website domains that are not monitored in our analysis. Since “YouTube” is the most popular video sharing domain, a video created in it will be linked and viewed by a large audience from a multitude of domains. This universal multimedia appeal is apparent in the Domain UoI variance as seen in the second and the third checks. Although Reddit users are interested in the post, their peak in the fourth check occurs later than the highest Domain UoI variance. Twitter is also providing a measurable variance in its UoI, with the highest mark at the fifth check. Interestingly, the Facebook UoI had a pretty high variance from the seventh to the fourteenth check, displaying the same late and persistent UoI variance, until the final iteration of the monitoring procedure, as seen also in Pattern 4.

This pattern describes the case of newly created videos in “YouTube” that are reachable by a multitude of OSNs and domains. They go viral within the first 2 hours of their creation, nettling the interest of Reddit and Twitter users before advancing to Facebook, with a long-lived exposure. As usual, the Domain UoI are accumulated rapidly in thousands, influencing the subsequent domain UoI variance. One hundred views are not of significant value, in a post with several thousand views. As before, ***Figure 31*** gives a great visualization of the spikes in interest, whilst **Figure 32** gives a perspective in the continuity of the variance differences for the entire monitoring process.

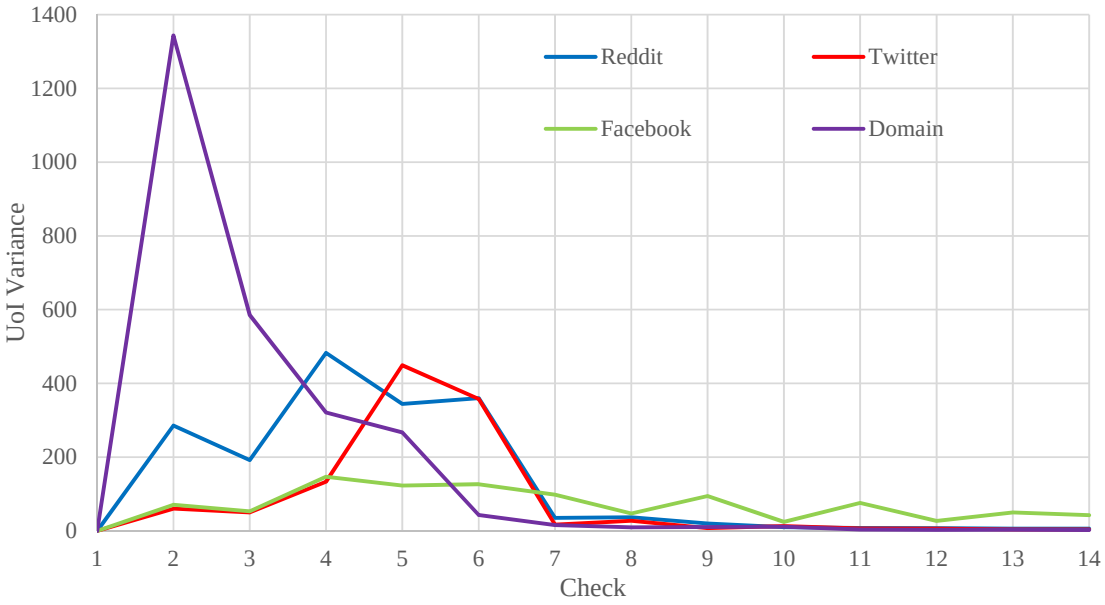

***Figure 31**. UoI variance per check*

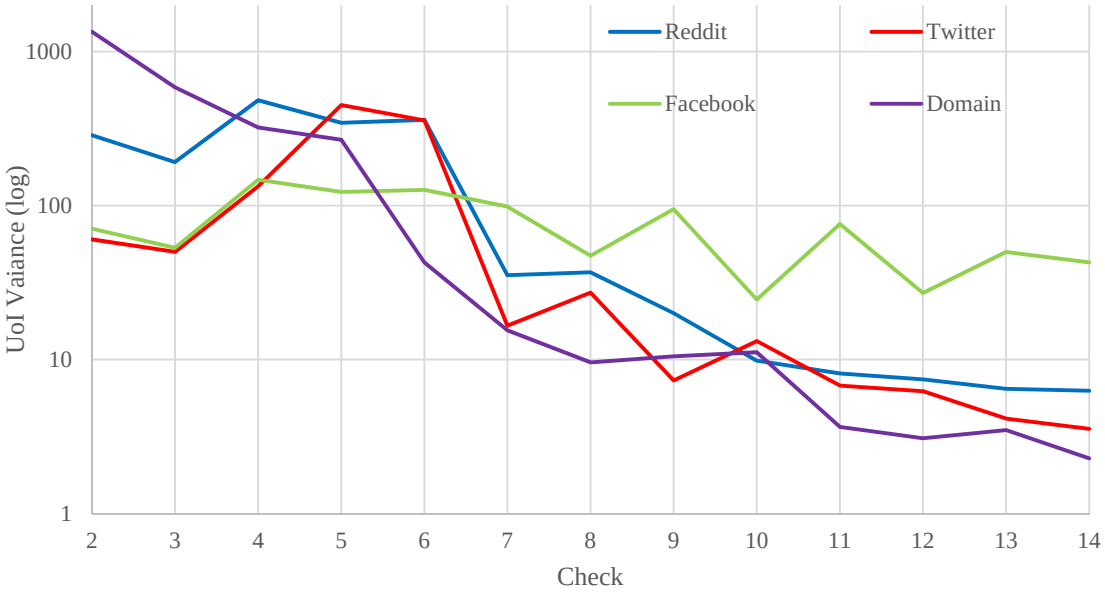

***Figure 32**. UoI variance per check, log scale*

3.3.6. INFORMATION DIFFUSION

The propagation analysis verified many observations of previous studies, mainly with respect to the micro and macro effects of information flows [1], [2], [30]. The

following happens when a post within Reddit is submitted. It slowly accumulates views, rarely going viral, up to a point where information is linked to various OSNs. Inside these OSNs, information is pushed from individual users to their respective social circles until the interest dies off. Out of the multitude of posts we encountered, post from topics with positive hue were found to be the most commonly shared content as mentioned in [7] and [10]. As observed in [14], the OSN content is more popular during the first few hours after its creation. However, this effect is amplified by the ranking algorithms of Reddit, with new posts being crucially designated –as viral or not– by a small number of early votes.

Furthermore, out of the total 550 posts that were shared in any OSN, more than 80% dealt with entertainment content, as mentioned in [9] and [10]. The most shared topics that comprised this type of content were; “AdviceAnimals”, “Funny”, “Gaming” and “Movies”. Emotive posts, as mentioned in [7], were also shared but not in the same degree as positive content. Nonetheless, due to the low number of such posts, no conclusions should be made for their OSN propagation. On the other hand, topics as “Gaming”, “Movies” and “Music” contained very specific entertainment information, which proved enough to provide a great diffused to submitted posts ratio. Lastly, “TIL” was the only informative topic and was found in rare occasions in both categories.

Based on the monitoring procedure described, two separate models for the information diffusion will be presented. Each model will describe the mean UoI and diffusion time. The first one will be based on content hosted in “ImgUr”, whereas the second one will be based on posts that linked to “YouTube”.

3.3.6.1. The ImgUr Diffusion Model

As seen in Patterns 1, 2 and 3, the ImgUr content appearing in Reddit, is mostly created from a user participating in this particular OSN. This affects the diffusion ties of

the hosting domain and Reddit. The information spreads in Twitter is slightly slower, followed by an even slower spread in Facebook. The UoI allocation is expected to be high for Domain and Reddit, and low for Twitter and Facebook, mostly because of the aforementioned content tie and the fact that the original post originated from Reddit. Apparently, if content was created in another OSN, then the same OSN would have a higher UoI than every other diffused OSN UoI.

The propagation time for every OSN is illustrated in ***Figure 33***, along with the respective UoI allotment. As mentioned, most of the Reddit posts with "ImgUr" content, are solely created for the purpose of the Reddit submission. Consequently, there is no delay in the propagation from the "ImgUr" Domain to Reddit. Let's us define t_0 as the time of creation in the "ImgUr" domain as well as the time of submission in Reddit. After 3 hours, links from Reddit or the hosting domain, start appearing thinly in Twitter. The time of transition of links in Twitter is defined as t_1 in ***Figure 33***. Almost 12 hours after the initial post creation noted as t_2 in ***Figure 33***, even fewer links appear in Facebook public posts.

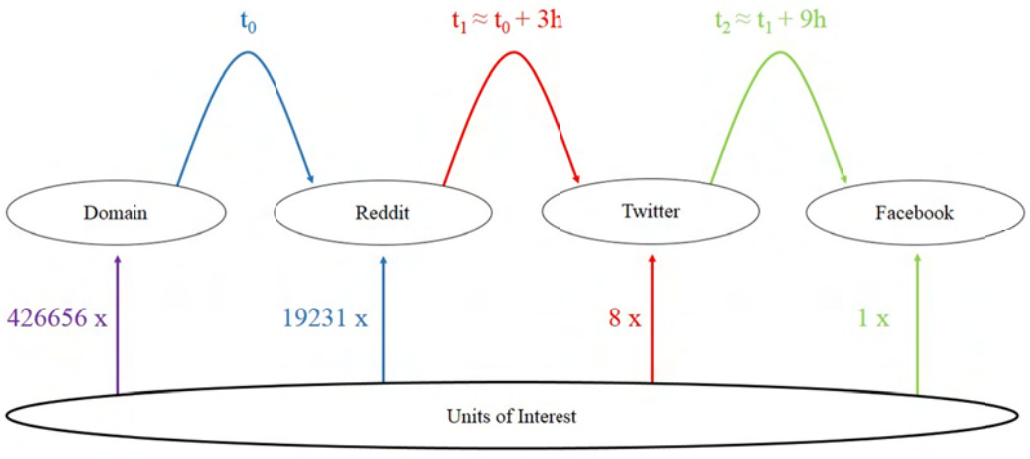

Figure 33. *Propagation Time and UoI allotment*

In addition to the propagation time for every OSN, the UoI allotment is also presented in ***Figure 33***. This UoI allotment is the direct ratio of the discovered UoI on Domain, Reddit and Twitter over Facebook. So a single UoI in Facebook corresponds to a UoI of 8 in Twitter, 19231 in Reddit and 426656 in ImgUr. Interestingly, the ImgUr over Reddit ratio is almost 22, fully descriptive of the extensive external linkage of "ImgUr" domain. Nevertheless, this analogy would be even lower, had we not used a strict UoI definition for each OSN. That would be because the UoI definition in "ImgUr" covers every domain view, but the respective UoI definitions for every OSN only cover references and interactions of the content. In *Section 3.3.6.3* we will further discuss one possible alteration in the UoI definitions and the resulting differences in the UoI allotment.

The mean numbers of UoI for the hosting domain and each OSN for every check are pictured in ***Table 21***. The disparity between each OSN compared to the domain is clearer, as well as the surge of interest in every domain. For the Domain UoI, the peak could be seen in the second check, followed by Reddit's most abrupt UoI growth in the third check. Twitter's UoI increase by 225% happens is found in the fourth check, with Facebook's most sizeable UoI increase –of almost 233%- seen in the sixth check.

Table 21. *Mean UoI per Domain*

Check (#)	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**
ImgUr	-	51509	65854	99575	127118	149445	172429
Reddit	-	374	1781	6430	10522	13114	15145
Twitter	-	2	4	9	13	16	19
Facebook	-	2	2	2	3	3	7
Check (#)	**8th**	**9th**	**10th**	**11th**	**12th**	**13th**	**14th**
ImgUr	189017	207424	223305	238900	250917	263924	273495
Reddit	16750	18136	19149	19971	20618	21213	21721
Twitter	21	24	25	27	27	28	29
Facebook	7	8	8	8	8	8	8

3.3.6.2. The YouTube Diffusion Model

Patterns 4 and 5 presented two distinct mean UoI variances observed in Reddit posts with linkage to “YouTube” content. However, most of that multimedia content was not specifically created for Reddit. Videos included in posts that passed the virality check were mostly created long before the submission in Reddit. In average, each video used in the discovered viral posts is more than 555 days old. So it was expected that Domain UoI would be significantly higher (in absolute numbers) compared to Reddit. Which in turn would have higher UoI count that our OSNs. As seen in ***Figure 34***, our expectations were confirmed.

A “YouTube” video was created 13352 hours –in average– before the Reddit post was submitted. Content appeared in Twitter after two hours, whilst the first public post of Facebook was seen three hours after the Reddit submission. Compared to “ImgUr”, multimedia “YouTube” content proved more pervasive both in terms of time and OSN propagation. After the initial submission in Reddit, “ImgUr” content required 12 hours to reach Facebook while “YouTube” content appeared in Facebook within 3 hours. One more indication of user interest in multimedia content throughout both OSNs.

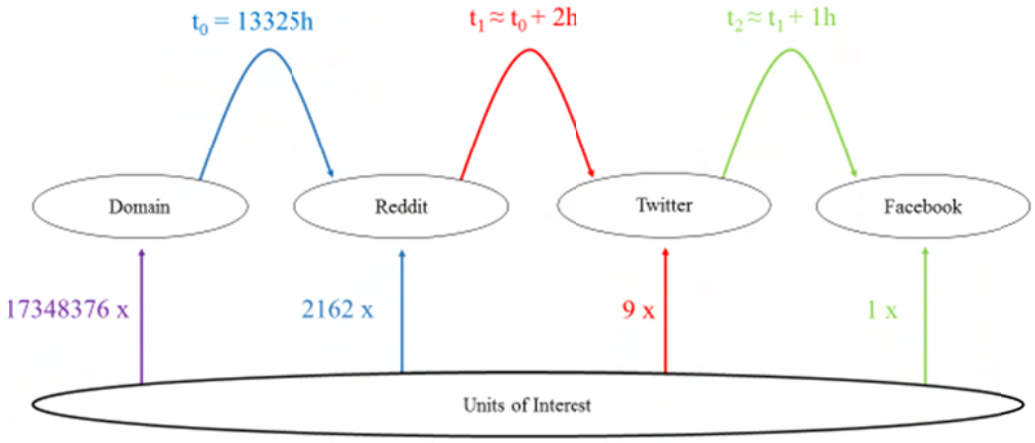

Figure 34. *Propagation Time and UoI allotment*

Of the total 52 “YouTube” videos, only 8 were posted in Reddit in the same day they were uploaded in “YouTube” domain. Overall, the domain virality period cannot be easily defined. Most videos already had a major number of views long before their submission in Reddit. After the first check, their mean Domain UoI variance is largely unaffected. Still, the ratio of either Twitter or Facebook UoI over Reddit UoI as seen in ***Figure 34*** is way lower than the one met in ***Figure 33***. Each Facebook UoI corresponds to 9 Twitter UoI, 2162 Reddit UoI and almost 2 million Domain UoI.

The mean UoI per check for every Domain are presented in ***Table 22***. As mentioned, Domain UoI interest spikes, in the fifth and the thirteenth checks, are almost unnoticeable. Reddit UoI steadily rise during the first 4 checks, probably due to the virality criterion, reaching their highest variance in the fifth check. Both Twitter and Facebook maintain a relative stability in the growth of their UoI. Additionally, their initial UoI are more than 10 times greater than those encountered in “ImgUr” (***Table 21***). The maximum UoI variances are: almost 22% for Domain in the thirteenth check, 989% for Reddit and 23% for Twitter both in fifth check, and only 7% for Facebook in the second check.

Table 22. *Mean UoI per Domain*

Check (#)	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**
YouTube	-	1197409	1197785	1329408	1546500	1566521	1586154
Reddit	-	28	97	305	3321	5554	7897
Twitter	-	269	296	327	368	454	505
Facebook	-	132	141	150	158	162	165
Check (#)	**8th**	**9th**	**10th**	**11th**	**12th**	**13th**	**14th**
YouTube	1607647	1613988	1643021	1669383	1680082	2048028	2070141
Reddit	9133	10489	11695	12785	13665	15389	16120
Twitter	544	594	637	665	696	730	755
Facebook	170	175	177	183	186	189	195

3.3.6.3. Extended UoI Measurement

For the purposes of a unified measurement, we defined UoI. In order for this measurement to be of a certain precision, we used several numerical counters in each domain and OSN. Although the inherit precision of a numeric counter was ensured, the only counters that precisely portrayed the traffic of the subjected domains were in "ImgUr" and "YouTube". In every OSN case, we used counters that could provide evidence of propagation. In Reddit we used the voting counters, in Twitter we counted the most recent (7-day period) mentions of the Reddit or the domain URL, in Facebook we also counted the mentions of those two URLs, but only on public posts due to API limitations, and in both domains, "ImgUr" and "YouTube", we used their respective view counters.

In "ImgUr" and "YouTube" the measurement of exact traffic is used to define UoI. For each OSN, various counters were used for the respective UoI definitions. The disparity in these definitions of UoI led to further lack of homogeneity. Since Domain UoI were defined as the vote counters, total viewership would certainly be higher mostly because not every user - in any of the OSNs- interacts with the content they view (likewise, for Twitter and Facebook UoI). Through available statistical measurements, it is possible to estimate viewership in each OSN submission based on our defined UoI. This approximation and the new UoI allotment will be presented in this subsection.

Based on the official about page of Reddit[40], unique visitors in September 2014 were over 170 million, generating a total of 6 billion page view. If we go back to the time of the crawling, September and August of 2013, the number of unique visitors was closer to 70 million with a monthly page view number of 5 billion. On average, between the crawling date and November 2014, Reddit had 106297343 unique visitors and

[40] http://www.reddit.com/about/, Last Retrieved: 13/01/2016

5170791876 page views. Considering that we are also provided with the mean number of votes per day, we can easily calculate the mean number of votes per month.

The daily average of votes during the crawling period was 17367892, while in November 2014 that number is slightly higher with a daily average of 21817065 votes. The mean number of daily votes for that interim period is 19592478, while the corresponding monthly average is 626286492 votes. By dividing the monthly votes to the monthly page views we calculate the vote participation on a monthly basis. During that 14-month period the monthly average of votes ranged from 9.58% to 14.12%. For August and September 2013, the monthly participation was 12.64% and 9.58% respectively. The monthly average votes value for the crawling period is 11.11%. This percentage is the ratio of page votes per views for posts found during the crawl period. For example, if a post in Reddit has 100 total votes, by using the monthly participation average we can estimate the corresponding viewers, which in this case would be 900.

In Twitter, every tweet is viewable by most –if not all– of their followers. That is the viewership range of a single tweet, and could be considered as the viewership of any original tweet. One of the few reliable and extensive studies in demographics and statistics of Twitter from Beevolve.com[41] showed that the mean number of followers for Twitter users is 208. A number that would be even larger if inactive and users with no followers were not factored in. Unfortunately, the data used for this analysis was of 2012. The increased usage of Twitter throughout the last 2 years would probably affect that average. If we apply the same rationale with the Reddit votes and views ratio, then we can multiply the number of Reddit and Domain URL mentions in Twitter with the mean number of 208 followers, in order to estimate Twitter viewership.

Table 23. *Mean estimated extended UoI per check*

[41] http://www.beevolve.com/twitter-statistics/, Last Retrieved: 13/01/2016

Check (#)	1st	2nd	3rd	4th	5th	6th	7th
ImgUr	-	51509	65854	99575	127118	149445	172429
Reddit	-	3366	16031	57876	94707	118038	136319
Twitter	-	416	832	1872	2704	3328	3952
Facebook	-	700	700	700	1050	1050	2450
Check (#)	**8th**	**9th**	**10th**	**11th**	**12th**	**13th**	**14th**
ImgUr	189017	207424	223305	238900	250917	263924	273495
Reddit	150765	163240	172358	179757	185581	190936	195509
Twitter	4368	4992	5200	5616	5616	5824	6032
Facebook	2450	2800	2800	2800	2800	2800	2800

In Facebook, one similar metric that could be used is the mean number of friends. Since every post -public or not- is viewable by any interested friend, the mean number of friends would expand the viewership public as well. The mean number of friends in Facebook -for adult users, is around 350[42,43]. As in Twitter, by multiplying the mean number of friends with the total Facebook mentions of a Reddit or Domain URL, we get a simple viewership estimate. We searched for a correlation between public and private posts in order to further improve that estimation but no such validated correlation exists.

If we take into consideration these viewership range estimates, then ***Table 21*** and ***Table 22*** should have their values recalculated. Nonetheless, these new recalculated tables should only be considered as rough estimates of viewership. The recalculated tables of extended UoI, ***Table 23*** and ***Table 24***, are presented below. Domain UoI are unaffected, since counters of ImgUr provide the actual viewership count.

Table 24. *Mean estimated extended UoI per check*

Check (#)	1st	2nd	3rd	4th	5th	6th	7th
YouTube	-	1197409	1197785	1329408	1546500	1566521	1586154
Reddit	-	252	873	2745	29892	49991	71080

[42] http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/, Last Retrieved: 13/01/2016

[43] http://www.statista.com/statistics/232499/americans-who-use-social-networking-sites-several-times-per-day/, Last Retrieved: 13/01/2016

Twitter	-	55952	61568	68016	76544	94432	105040
Facebook	-	46200	49350	52500	55300	56700	57750
Check (#)	**8th**	**9th**	**10th**	**11th**	**12th**	**13th**	**14th**
YouTube	1607647	1613988	1643021	1669383	1680082	2048028	2070141
Reddit	82205	94410	105266	115077	122997	138515	145095
Twitter	113152	123552	132496	138320	144768	151840	157040
Facebook	59500	61250	61950	64050	65100	66150	68250

UoI allotment is heavily affected upon considering the extended UoI definitions. The results portrayed in ***Figure 33*** and ***Figure* 34**, bear only a small resemblance to the new UoI distribution. ***Figure 35*** shows the inextricable connection of an "ImgUr" upload and a Reddit post, while ***Figure 36*** presents the inequalities formed by using old content within Reddit. The connection is seen in ***Table 23***, where "ImgUr" and Reddit UoI are almost identical. ***Table 24*** presents the aforementioned UoI discrepancy of "Youtube" and Reddit UoI.

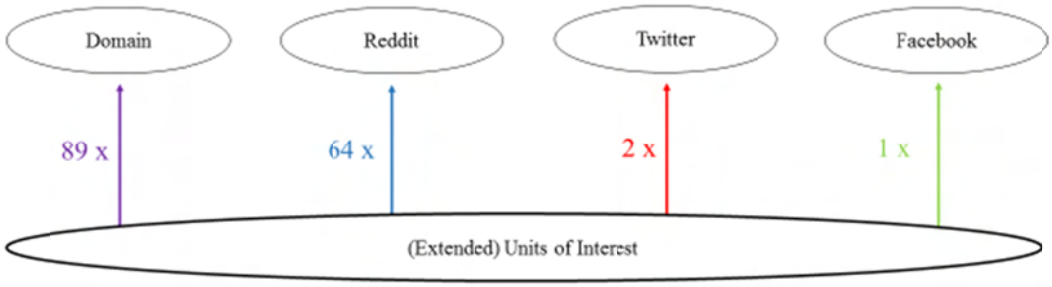

Figure 35. *Extended UoI Allotment, "ImgUr"*

Figure 36. *Extended UoI Allotment, "YouTube"*

In this section we presented an estimated analysis of information propagation. It should not undermine the study based on numerical counters, which is a factual analysis, but should rather complement it. This estimation approach also abridges the gap found in Facebook UoI analysis, which could be further improved if there existed a detailed correlation of private and public posts. Concluding, the extension of the UoI definition should not impact the approximate propagation timeframes based on the observation of social interaction's "Ripple effect" [140]. "Ripple effect" states that social interactions act like droplets in still water creating expanding outward ripples incrementally in relatively short timeframes.

3.3.7. DISCUSSION

Our research is the first that considered multi-OSN propagation and the layered perception of information flow, which will be presented below. In addition, some interesting results were derived concerning information diffusion in OSNs. On top of all, several propagation patterns were observed for content originating, linked and appearing in Reddit. Viral content gets a high enough level of viewership within the posting OSN, before users start conveying it to other OSNs. Upon the transition to other OSNs, Reddit interest increase is slowing down. Posts with positive or entertainment content were the most viral. Content persistence was relatively low with interest fading within the initial 24 hours. Exceptions were found in "Gaming" posts, and most of them had at least some level of interests well past the 24-hour mark. Lastly, "Movies" posts were the only ones that propagated almost simultaneously in every OSN.

Based on the general propagation behaviour of Reddit content, we present a new perception of information flow. One which would accurately model the multi-OSN

diffusion. In order to provide a simple example of that propagation perception, let us use a random Reddit post. A post that follows the diffusion Pattern No. 1 shown in ***Figure 23*** with 30 Domain UoI, 20 Reddit UoI, 7 UoI in Twitter and 1 in Facebook. Our example post has just propagated to the last OSN, Facebook.

Information is diffused between users and multiple OSNs. A respective figure, with random placement of users, would probably be similar to ***Figure 37***. The Domain content is viewed by multiple users (purple nodes), while Reddit initial post is mostly viewed by users within the OSN (blue nodes). Information then propagates to Twitter and Facebook (red and green node respectively). The randomness is not descriptive of the propagation order, nor the distinct domain and OSNs.

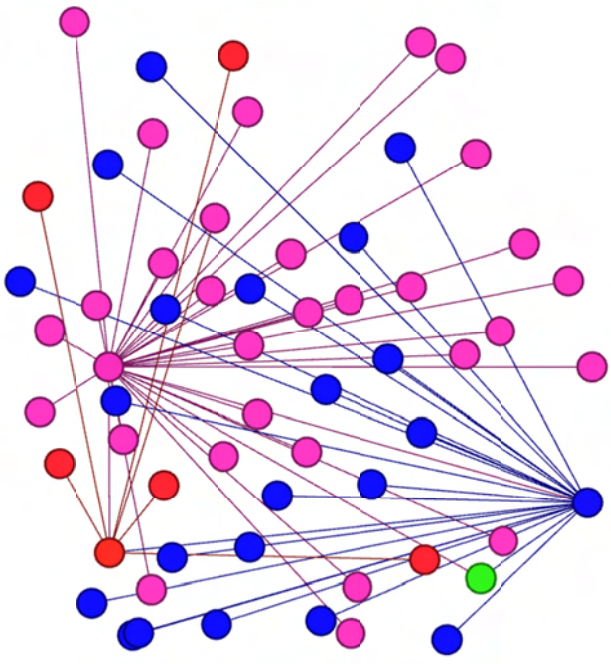

Figure 37. *Random Diffusion Visualization*

By performing two simple steps, the layered structure of information propagation can be distinguished as presented in ***Figure 38***. The first step, in order to achieve such a

clear representation, is to distinguish the respective layers based on the nodes with the highest degree for each layer. This way we ensure that when a weighting clustering algorithm is applied, the clusters -which would be formed around the highest degree nodes- will follow the propagation order of the post. The second step is to apply any clustering algorithm -weighted or not- and observe the distinct domains formations. In our example these groups can be seen in the following top to bottom order: Domain, Reddit, Twitter and Facebook. The clustering algorithm we applied was “Forced Atlas 2” [141].

Purple nodes denote the Domain users at the top, Reddit users are drawn in blue below domain ones, while Twitter and Facebook follow next – at that order – with red and green nodes towards the bottom. This lateral propagation is portrayed in relation to time of diffusion, and the lateral diffusion is happening concurrently with the horizontal expansion. As interest is shared between OSNs, each network enjoys its own interest propagation. After approximately twenty-four hours, diffusion and graph expansion halts.

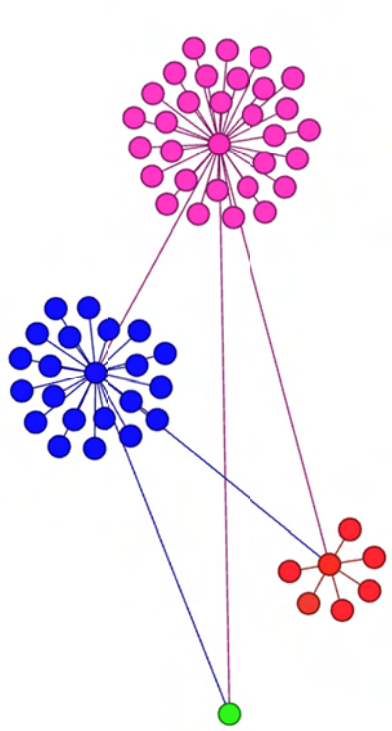

***Figure 38**. Layered-based Propagation Visualization*

4. AN ANONYMOUS ONLINE SOCIAL NETWORK OF OPINIONS

4.1. Online Social Parity

Apart from the structure and content properties, OSNs issues are considered such as the social impact and possible implications. They are thoroughly studied in various online network environments that were presented in previous sections. In addition, most of the effects and peculiarities of OSNs were initially observed in offline social networks. Therefore, by analysing older social network behaviours and studies OSNs could be further studied, improved, or even get altered. Thus, we propose a new form of OSN without data access limitations that would aim towards productivity and bias elimination.

4.1.1. SOCIAL ISSUES

Social interactions and networks have long been in the centre of sociologists' interest. From the early 1940s, studies have been made around social interactions and their subsequent networks. Most of these studies, were focused on small communities such as families or a particular group of individuals [141], [142], [143] [144] and [145]. In few of them, authors tried to detect shared patterns and properties amongst every social network [146], [147] and [148]. Unfortunately, literature on social interactions for networks not bound by personal relationships is pretty limited. As it was impossible for two unknown -to each other- individuals to communicate in a traditional offline network. However, social structures and norms were also studied regardless of the underlying network structure [149].

One of the earliest studies in social structures was [146]. Radcliffe-Brown described the social phenomena in human societies as immediate results of the social structure by which humans are unite with social relations being a part of the whole social relations network and the social structure. He also differentiated individuals and classes

based on their social role mentioning “The differential social positions of men and women, of chiefs and commoners, of employers and employees, are just as much determinants of social relations as belonging to different clans or different nations”. He further underlines the dynamic nature of social structure with relation links readjusted constantly.

One year later, Moreno introduced the foundations of Sociometry [148]. The author described the chief concepts and presents the most significant discoveries of sociometric studies. An interesting notion was that of the catalyser which is described -on a social level- as the spontaneity of all the individuals in the given society. He also mentions that “Sociometry opened up a new possibility of genuine planning of human society for the reason that the factors of spontaneity, the initiative and the momentary grasp of the individuals concerned were made the essence of the method of exploration and of the investigation itself.” In sociometric systems, total spontaneity is the essence of planning rather than the spontaneity of a small number of individuals or leaders.

In [149] social power in human society was discussed. Political science, sociology and social psychology are some of the society related field where social power distinctions can be observed. The aim of Raven was to identify and define -in terms of influence- these distinct types of social power. The kind of power that can lead to alterations in behaviour, opinions, attitudes, goals, needs and any other personal psychological aspect. They also noted the necessity for the existence of a social agent in order to influence an individual of the respective social system. The crucial factor in that influence relation is the relationship of the agent and the individual. Ultimately, five types of power were defined with notable remarks. The stronger the relationship of the agent and the individual the greater the power, and thus the influence. Out of all the discovered forms of power, referencing power had the broadest range. Coercion and

reward powers had opposite results. Coercion decreased relationship attraction and presented high resistance, while reward power increased relationship attraction and lowered resistance. The negative effects of coercive power were disproportionate to the legitimacy of the coercion.

A re-evaluation of the social notions of survival and exploitation is presented in [150], based on the concept of reciprocity. Gouldner also analysed reciprocity but separates:

- Reciprocity as a pattern of mutually contingent exchange of gratifications
- The existential or folk belief in reciprocity
- The generalized moral norm of reciprocity

Although the study's character resembled a personal essay, it provided an insight on reciprocity's contribution to social stability.

Bert N. Adams in [151], outlined the consensus and positive concern variables of social relationships along with their implications in the social network. They distinguished social structure in two divisions, kin and non kin. He also mentioned various subdivisions of kinship and different definitions of social distinctions and interactions. Finally, various propositions of social interaction were presented based on the broader type of relationships such as kinship and friendship.

The network approach of urban studies was introduced in the 1970s and differentiated itself from any other approaches of social analyses. Craven and Wellman [152] focused on who was linked to whom and the resulting structure of the relational network, while addressing specific problems of urban network studies, such as structural ties, migration, resource allocation, neighbourhood and communities. The notion of a network of networks was once again brought up in [153]. Wellman noted that the scale and diversity of an urban network were the source of strength rather than of chaos, and

the provision of efficient means for personal communication within similar networks was mandatory.

In [154] the accuracy of the informant within social networks was addressed. Killworth and Bernard distinguished the four considerations of error assessment of sociometric data collection: validity, reliability, veridicality and accuracy. They described each consideration and presented a common problem around it. They also collected network data in order to assess the identity and quality of the informant based on the communications of the networks members.

In organizational media use, Fulk et al. [155] studied the significance of social influence processes and social attitudes and behaviours. These were determined partially by the information of the social context as well as the observations of an individual's past behaviours. The social model provided explained a wider range of existing findings and could be used in design, conduct and report for social environments within organizations.

In the early 1990s studies for computer networks were mostly -if not only- focused in organizational networks. [156] starts with a really insightful note "The image of computers being used to empower people is a potent and appealing one". Clement also identified power as a dependency of relations rather than a personal matter, since people mostly functioned in communities. Collaborative informal networks were also discussed, with the overall research asserting the service of human needs will be the judge of computer revolution rather than computer complexity or power as mentioned in [157].

At the same time Wellman and Wortley [158] studied the way different ties provided different supportive effects on networked individuals. They recognize six kinds of supportive resources: tie strength, contact, group processes, kinship, characteristics and comparison of network members. At the same time, they explained the social production of social support. The research is based in a small sample of an urban area in Canada,

which on the one hand grants a sense of homogeneity of the network, but in the other hand it is was based on a very small sample. They concluded that networks in which certain social aspects are resolved, the supportive relations are smoother compared to networks where insecurities (e.g. political or economic) exist.

Similarly, social network structural analyses was researched in [159], and five characteristics of the underlying intellectual unity were distinguished. These were:

1. Behaviour as interpreted in terms of structural constrains on activity
2. Analyses focus of relations between individuals
3. The patterned relationships and their jointly affection of members' behaviour
4. The network of networks may or may not be partitioned into discrete groups
5. Analytical methods deal directly with the patterned, relational nature of social structure

For the first characteristic, it was noted that the individualistic nature of members forced an unintentional deterministic behaviour towards a certain goal. While for the second feature, it was argued that relational analyses were preferred over similar categorical analysis methods which sorted members based by their attributes.

A digital money transfer system was proposed in [160]. A system that could be easily modified for information exchange. In that context, the authors outlined the need for a reputation and incentives system for customers and sites. In addition, Hardy and Tribble highly valued the need of security and reliability. Concluding they noted that the most crucial element in digital money exchange scenarios was anonymity. If money exchange is paralleled with information exchange, do the same notions apply? However,

information can only be considered as a source with subjective -rather than exchange-value.

Ethics is one of the rarest subjects in social networks studies. In [161] Mackay addressed the ethics behind the use of multimedia capturing in human computer interaction scenarios. Once again, the subject is slightly off from strict social networking but the general remarks can be applied in any monitored social environment. She analysed the use and misuse of monitoring, its ethical confines, and provided some general guidelines. Privacy and the need for protection of users were the closing arguments.

In a much later study [162], the computerization of our society and its effects in social structure were addressed. Burkhardt and Brass noted that early adopters of new technologies had a greater increase in power and centrality compared to late adopters. Lining structural position in the established network as a factor for power redistribution. They also provided real life experiment results and analysis based on environments that underwent significant structural changes.

In [16], which was already discussed in *Section 1.2.1*, Haythornthwaite drew attention to the empirical observations and along with relationships formed the basis of social network analysis. It was also noted that network position is affected by who controls, facilitates, or inhibits the flow of information. The need for unbiased assumptions over information exchange among peers was highlighted in order to observe where information exchanges were happening. Concluding, the reasons for examining social structures were the identification of information types, the individuals in control and the way information flows from one network to another.

Interestingly, one of the first studies about anonymity in social groups [163] was published in 2001. In particular, the authors researched communication effects amongst

anonymous and identifiable computer communication groups based on established studies. They suggested that, in a group with separate levels of identification anonymity further enhances group's salience. They closed their essay with the following phrase: "Paradoxically, reducing the presence of the individuals within the group may actually serve to accentuate the presence of the group within the individual."

Remarkably, O'Reilly in [69] revealed the lack of experimental studies in new OSNs. An observation that coupled with the limited understanding of why and for what purposes users utilized OSNs in their early days, is still applicable –though in a smaller scale.

There also exist several books that were based in similar observational environments of social networks. As mentioned, early remarks could potentially be applied to the corresponding online counterparts. The first and most renowned is [164] of 1979, with more than 20 researchers and 25 chapters. Almost each author contributed a separate chapter and the vast majority of theoretical social network concepts were presented.

During the internet era, various authors have attempted to imprint the basis and analyse social network behaviours and norms [165] [166]. In the first book, the most recent methods of analysing, sampling and measuring of social networks were presented complementing an earlier work of 1995. While in the second one, an overview of the scientific community research stance was presented in relation to the newly forming online communities.

Most of these studies and books were dealing more with social environments in their traditional and general sense rather than exclusively online networks. Additionally, most of the remarks and results of these studies are not affected or characterized by the social identity of the network, nor the physicality or its absence of relations.

4.1.2. Online Social Issues

As shown, modern OSNs have many similarities. Most of them have a profile page, let you connect with other users, evaluate and share any type of content etc. Even if one of the above characteristics is missing, in a specific platform, the rest outweigh its absence. For example, in networks where connection and ties have a secondary role, the sharing element is usually way more prominent.

However, in almost all of them, users tend to behave in a strange but expected way. Their behaviour is expected, because online users are projections of each one's personality, with its benefits and drawbacks. While the peculiarity lies in the strange use of every free information online network. The term peculiar is employed due the fact that our offline social life was ported online, not only once but in multiple occasions, almost identical to traditional social networks.

From the euphoria and directness of social communication through online social media to the serious criticism for most of the OSNs. Let us examine some of the most evident features of online communities, features that are not inherently negative but tend to have negative results.

4.1.2.1. Privacy

Online privacy covers the right of a user to not reveal personal or any other type of sensitive information, regarding the use or interaction in online environments. A right that is mostly overseen by users and poorly preserved by most OSNs. Most concerns are raised in relation to disclosing personal information for advertising purposes. However, in several cases, OSNs have used personal data directly for profit[44].

[44] http://www.reuters.com/article/2012/03/01/twitter-data-idUSL2E8DTEK420120301, Last Retrieved: 13/01/2016

The scientific literature on privacy in OSNs is both rich and substantial. From the era of MySpace and early Facebook Dwyer et al. [167] raised the concerns of users' privacy within these networks. They also dissected the members of the networks based on they behaviour. Facebook users were found more prone to reveal personal information, while MySpace users were developing more personal relationships. Facebook is inextricably linked with privacy concerns [168] [169] [170] [171]. In Twitter, research around privacy is not so sharp mainly due to the nature of the OSN where personal information is not mandatory. However, there has been multiple situations in which users reveal sensitive information either unwillingly or naively [172] [173]. A very interesting solution, was proposed in [174] as a secure alternative, in order to protect user messages from unwanted eyes.

Most of our concerns in regards to data, its access or usage, were presented in the previous chapters. However, there has been cases, outside of the research spectrum, that raised some concerns, most recently Facebook used 700000 users in order to study emotional contagion without consent from the users[45]. In Twitter, apart from the case of data sale we mentioned, the network has been profiting from selling data access to several data management companies[46].

4.1.2.2. Tightly Knit Communities

Online social communities also display the effect of tightly (close/closely/dense/densely) knit communities. This effect describes the clustering of people based on common interest, inveteracy, location, history or any form of social tie. As with most OSN characteristics, it was first mentioned in various offline networks

[45] http://time.com/2949565/heres-what-facebook-can-do-with-your-personal-data-in-the-name-of-science/, Last Retrieved: 13/01/2016
[46] http://business.time.com/2013/10/08/twitter-is-selling-access-to-your-tweets-for-millions/, Last Retrieved: 13/01/2016

studies to describe small communities, where members of the community would interact frequently with each other. It was later observed in multiple online communities with various effects.

Most notably, Chau and Xu observed in [175] that densely knit communities and the community interaction patterns of a greater network had important implications to the functioning of the network. They also noted that the tightly knit community became the collective identity of its members, and that personal nature is usually diminished within such networks. While in [176], Cummings et al. pondered whether closely knit communities, where people developed a sense of belonging, was the norm. To counter that theory, they analysed a number of group email communications, where indeed the nature of the relationship proved to be voluntary rather than collective imposing or consensus. Lastly, closely knit communities as a source of social capital was identified in [177]. Social capital is defined as the collective benefit of the network and its members through the effects of interest and goal commonality.

This type of behaviour has multiple effects in online networks, but mainly excludes new users and polarizes existing users. A great example is Wikipedia [178], where editors are forming two distinct group with each group re-editing and correcting the other groups edits. These communities also enhance the effect of diffusion of responsibility, making them more unconcerned for their online behaviour [47] [179], [180].

4.1.2.3. Power Law Distribution

Power law is a relationship that links two network properties based on a comparative connection. For example, in most social networks users and linkage are characterised by a certain power law relationship. This power law distribution was

[47] http://www.newyorker.com/tech/elements/the-psychology-of-online-comments, Last Retrieved: 13/01/2016

evident in offline networks but reached to a new scale with the advent of OSNs. From the earlier web [181] [182] and blog sites[48], online forums [183], to most modern OSNs [48]. Many variations of Power Law type distributions exist. Of them, the best known is Pareto Principal [79] which was extensively covered in *Chapter 2*.

In one of the first studies based on OSNs [80], Adamic et al. researched Gnutella, a peer to peer network. They exploited its power law distribution in order to create efficient search algorithms. In [184] Sala et al. tested estimation algorithms for power law analysis in Facebook, concluding that -for their sample- the estimation for the power-node group was smaller than expected. Prof. Krugman also noted[49], Twitter users have an almost power law distribution[50]. A. L. Barabási and Z.N. Oltvai [185] observed that when new nodes connected to a network, they tended to connect to an already strong node, thus creating (non-random) connectivity power law distributions.

The greatest drawback of power law distribution in social networks is information visibility which leads to opinion shaping. As mentioned, the two step flow of communication [109] is evident in this type of networks. In traditional OSNs, a post from a specific user is initially visible to everyone who is connected to that user. The immediate effect is that users with fewer ties encounter a weak propagation of their information even if their posts are more entertaining, meaningful or constructive. A more indirect outcome of this linkage is the information "suppression" that new users experience, and upon connecting to an already established network, time is required in order to build connections and improve your profiles' range.

[48] http://shirky.com/writings/powerlaw_weblog.html, Last Retrieved: 13/01/2016
[49] http://krugman.blogs.nytimes.com/2012/02/08/the-power-law-of-twitter/, Last Retrieved: 13/01/2016
[50] http://blog.luminoso.com/2012/02/09/twitter-followers-do-not-obey-a-power-law-or-paul-krugman-is-wrong/, Last Retrieved: 13/01/2016

4.1.2.4. Bias

Bias was one of those social features that not only was transferred to the online communities but also evolved of shorts. Bias is the behaviour in which an individual expresses or forms a certain view towards a group or another individual without having prior knowledge or simply based on preconceived notions (stereotyped or not). Similarly, in online environments users tend to be biased towards certain individuals or communities, while the positivity or negativity of their bias is irrelevant.

Bias aspects in OSNs are largely unexplored. Based on limited sociology studies, some insight into online bias could be provided. A very interesting study on Persuasion Bias [186] described how individuals and groups were influenced. DeMarzo et al. noted that influence on group opinions not only depended on information accuracy but also depended on how well connected a given user was connected to the social network. They also dealt with the phenomenon of unidimensional opinions, with opinions over versatile issues converging to a single “left-right” spectrum.

Lin et al. [187] noted that in modern OSNs “There are more voices than ever, but many are echoes”. Underlying the effect of opinion reproduction instead of individual opinion forming. Counts [189] concluded that users in Twitter are biased positively towards their friends and negatively to unknown users. He also noted that the name of a user affects our perception with respect to the information the user shares. Last but not least, social media bias in relation to location specific weather events were studied in [189]. Kiciman outlined the complexity of social bias, as well as their connection with sentimental analysis.

4.1.2.5. Credibility

Credibility is the term that describes whether a person is trustworthy, in general or in a specific subject. Usually it is based on expertise or proven knowledge on a subject.

Online credibility is widely studied topic since the earlier days of the Internet, it is a greatly controversial issue mainly because counter arguments -coming from weak nodes- are rarely getting enough visibility, while strong nodes of these networks rarely have to prove their expertise or knowledge.

Nowak and Rauh [190] confirmed a process of uncertainty reduction through interaction, in particular via still images. Westerman et al. [191] collated the perspective of individual twitter users that inclined to attribute expertise, based on the followers/following ratio. They also pointed out that relaying information speed is counterproductive towards judging a Twitter user's credibility. Metzger et al. [192] suggested a group based credibility assessment in online networks, trying to enhance the heuristic process of assessment. Recently, credibility in a travel rating site was studied based on user-generated content [193]. Ayeh et al. concluded that "perceptual homophily" was crucial in credibility assessment.

A most recent case of credibility in OSN is that of a prolific user in Reddit[51]. This particular case outlined general problems in relation to rewards systems. In that case, one user was exploiting the Reddit vote system in order to be seen as the most credible and popular commentator in biology related issues. In relation to other OSNs, Facebook is filled with fake accounts[52] and microblogging site Twitter's credibility assessment is complex and time sensitive [194] [195].

4.1.2.6. Information Quality

Quality assessment of online information is a rather old problem [196] [197] and OSNs provide the most challenging modern application. Various proposals have been

[51] http://www.dailydot.com/news/reddit-unidan-shadownban-vote-manipulation-ben-eisenkop/, Last Retrieved: 13/01/2016
[52] http://www.sec.gov/Archives/edgar/data/1326801/000119312512325997/d371464d10q.htm, Last Retrieved: 13/01/2016

made concerning the segmentation and analysis of the information sources [198] [199], as well as the information itself [200] [201]. In contrast to these previously proposed techniques OSN information analysis should also take into account issues that were not critical in pre-existing information media. These issues include the direct interaction between users and information media, user participation, collaboration, information sharing, as well as timeliness and context [202] [203].

4.1.2.7. Various Issues

Moderation issues are also prevalent in various online communities, more commonly in forums and such. The difficulty of evaluating a moderator's work is highlighted[53] and their role in promoting free political speech is questioned in [204]. Ratings and ranking mechanisms in OSN environments force users in posting pretentious content in order to boost their profile visibility. In addition, users spend time to maintain their online profile, filter their interactions, social ties and sharing. This excessive time spent in online social networks can be seen as addiction based on various psychology studies [205] [206] [207].

All of these issues, along with most of human traits and flaws are dominant in OSNs. As is elegantly put in "The Psychology of Online Comments"[54] with the phrase, "The medium may change, but people do not".

What would happen if one didn't project her/his entire personality and interests in these networks? What if we could form a social network where a user is represented only by his opinion? Would we observe the same issues or the same interaction patterns, as in traditional connection based OSNs or they would be slightly altered? Finally, what would

[53] http://labs.yahoo.com/publication/who-moderates-the-moderators-crowdsourcing-abuse-detection-in-user-generated-content/, Last Retrieved: 13/01/2016
[54] http://www.newyorker.com/tech/elements/the-psychology-of-online-comments, Last Retrieved: 13/01/2016

be the underlying structure of its respective social graph? Even if it is not commercially feasible, we strongly believe that it be interesting from a research perspective.

4.2. The Online Social Experiment

4.2.1. Equality, Privacy and Data

Based in the aforementioned issues and peculiarities. We propose a concept of an online social network, one that aims in addressing most of the issues mentioned in the previous section. Moreover, its purpose will primarily be experimental and data will be free to access for research purposes. Although most of the concepts we will present have been in used in at least one Online Social Network, they have never been combined into one single network. It might even fail in the completion of our predetermined goals (such as the elimination of bias) but its success will be measured in scientific insight for various unexplored subjects rather than its impact or public's acceptance.

It is crucial for this new type of online network, to promote discussion, equality and productivity of any form. Whether this network will eliminate biases, generated by the usual connected structure of the social networks, is uncertain. Of outmost importance is the unobstructed access to its data, for research and non-commercial purposes. With great respect to the privacy of users and their data.

Therefore, the proposed platform will aim to:

- Counter most of the online issues
- Equally promote discussions
- Provide unrestricted access to network data
- Protect users' privacy

4.2.2. Design

Our first goal is to eliminate the majority of online issues as presented in *Section 4.1.2*. In the framework of our experimental network most of these issues will be directly addressed, while the remaining may require the assistance of scientists from various

disciplines. The most important issues that pertain our scientific knowledge and experience along with their proposed solutions are presented below.

Our first step towards equality is to cut familiarity ties that are portrayed in OSNs in the form of connections. Unfortunately, this would mean that users who used ties on a proper way will also lose that ability. However, the aim it to set a barrier to any biases generated from acquaintance by not having visible-to-users ties. In the same direction, usernames will be hidden so that posts are not categorized or judged based on identity. Furthermore, there will be no promotion of popular or highly rated posts that usually attract “undeserved” attention.

Equality on judgment is the next key aspect of the proposed OSN. Usually, content posted on an OSN is judged and rated by popularity. A widely positive or negative post often drives preference. By avoiding visible counters either positive or negative, we take a step towards isolating personal preference-judgement and public opinion forcing. However, that does not mean that voting or usernames won’t be monitored only that the voting/ranking/rating information won’t be visible to users.

What will be the structure of the network? How will the public information shape the social graph? Votes which resemble positive and negative affiliation, with respect to the author of a post and will provide a measurement of linkage. Every post vote will be strengthening or weakening the relationship of the voter and the user who made the post. Each post will fall under one broad category (e.g. sports, politics) and will be classified based on a set of keywords, similar to keywords in stackexchange, or hashtags in Twitter.

In order to promote discussion, as well as collecting information regarding users’ linkage and connections, there will be the ability to reply and comment in each post. Each reply will form a new discussion thread. By using the same principal as in the initial post, every reply and vote denotes (and will be monitored as) a possible affiliation link.

Finally, every user will have a merit in the evaluation of content, via implicit or explicit feedback.

4.2.3. Aim and Necessity

Above all, research interest is what pushes us towards a new concept of OSN. Interest for the analysis of interpersonal online relations, as well of the subsequent underlying graph data. While at the same time, promoting equality on both, opinion expressing and viewability. Last but not least, data access will be unrestricted to interested researchers. Apparently, the network should enjoy a certain level of adoption in order to spark researchers' interest.

Although that level of user adoption and the subsequent success, is not accurately defined nor analysed in depth by the scientific community, we strongly believe that our proposed OSN will aid in the analysis of the aforementioned unexplored properties. In addition, since productivity is a pretty undermined issue in modern OSNs, our aim is to create a form of productive (in any sense) anonymous online network similar to stackexchange and its derivative sites[55].

But, what would be the contribution of such a network in the already crowed OSN space? Do we need a network that almost disregards the personal element of a user? To these, and any similar questions, we quote "Everyone shall have the right to freedom of expression"[56]. We aim to uniformingly and equally promote opinions in a network that would be solely focused in them. The necessity of such a network may be subjective, but the creation or even the investigative concept of one is sufficient for a researcher.

[55] http://stackexchange.com/sites#, Last Retrieved: 13/01/2016
[56] http://www.article19.org/pages/en/international-guarantee.html, Last Retrieved: 13/01/2016

5. CONTRIBUTIONS

5.1. Contributions

Our original contributions to knowledge are the introductions of a new sampling method for OSNs, a novel multi-layered OSN diffusion theory and the concept of an innovative OSN. It is apparent that throughout our research we focused on the broader field of OSNs. Every problem, every solution, proposal or topic, led us to a closely related research. The research potential around OSNs is almost limitless since their scope, size and type change dynamically and unpredictably.

During the early stages of our research, we had to deal with the data access restrictions of Twitter. Throughout its sampling, we tried to discover ways of countering these restrictions and came up with an efficient sampling method that takes into consideration the limitations of the environment. Although our algorithm was created with Twitter in mind, it can be easily applied to various real life or online environments sampling scenarios. Our goal was clear, to introduce efficient data sampling for networks with data access restrictions.

Continuing our research in OSNs, we noticed the lack of inter-network diffusion studies. It is very common, for a post of a certain OSN, to be shared in multiple other OSNs. However, there were no respective study to quantify the diffusion impact or even lifetime of the shared content. In that regard, we studied how a post hosted in Reddit diffuses within the network itself, as well as in Twitter and Facebook. Every analysed post was either a picture hosted in ImgUr or a video hosted in YouTube. In this research, we aimed at studying the lifespan and properties of a post originating from Reddit and shared to other OSNs.

While in the third part of our studies, we focused in a more socio-centric issues. Upon studying successful OSNs, we wondered whether the absence of a personal profile

would affect user linkage. We proceed to propose a new type of OSN. based on the respective (but limited literature) on the social issues we identified in modern social platforms. Our ultimate goal is to create an experimental online platform that will provide an ideal environment for interested researchers, in order to enrich studies around online bias, equality and anonymity, as well as OSN themselves.

5.2. Future Research

We introduced a new sampling approach, applicable in online networks sampling situations. In addition, since every online social network is constantly evolving, the vast underlying graph is becoming even more complex and requires sophisticated sampling methods. Methods that should take into account rules on data access policies. Unfortunately, more and more OSNs appoint third parties to handle their data. This data handling model results in more limits on data access, several new modifications in metadata information and permission obstructions. Our proposed sampling enhancements could be seen as tools to overcome existing and potential limitations, in regard to graph data access and analysis. Sampling and crawling processes grant access to an almost infinite amount of information. Social network ties contain more information than the apparent entity linkage. Information analysis topics such as trends, sentiment and polarity are also bound to the underlying graph of the network. Thus, future applications of our proposed algorithm may not be bound to strict sampling.

However, several issues remained unanswered. For example, how accurate would such a sampling method be in “undirected” OSNs? How would an even broader selective distribution percentages work? How could we further improve the estimation of the degree distribution? Would these proposed enhancements work as effectively in other sampling methods? Every question could possibly lead to a future research subject.

Multiple OSNs analysis and testing of the proposed enhancements are the most prominent goals. It will be achieved by applying the same techniques on same sized samples in every possible OSN. Extending the thought and implementation about every type of network, we could test our sampling method in several networks and graphs even if they do not belong in the class of OSNs. A direction we took when we tested (in a small scale) the SGs in *Section 2.3.7.3*.

On our research in multi-layered information propagation, several improvements can be still made in order to better determine the overall diffusion speed and depth. Most notably, in order to assess the diffusion speed with accuracy time stamps should be employed throughout the crawling process. Furthermore, the monitoring algorithms should be focused on speed in such a way that virality would be detected as fast as possible. Despite the fact that many OSNs impose access limitations, see *Chapter 2*, a balance should be found between these limits and the accurate and timely identification of virality. In addition, the topic selection could be even broader, based not only on certain category criteria, but also on some form of Natural Language Processing. Although our OSN corpus was restricted, it was focused on the most popular OSN. The multitude of available OSNs is tremendous and allows for the expansion of the corpus in future studies with emphasis in virality, information flows and time.

Our future goal is to create a virality diffusion analysis tool, which will work in real time. One that could analyse a multitude of OSNs at the same time, we also hope that future OSNs will be open to scientific and non-profit research, and not behind closed paywalls. The applications of a real time analysis on viral posts are innumerable. The scale and usage of modern OSNs are the warranties for the plethora of research implementations. We also plan to introduce a new isometric visualization algorithm (for Gephi platform) that will automate the clustering process.

Further improvements can be made in several aspects of the current research. Interestingly, our analysis of Reddit, pointed to the power of collective action. Most notably in news reporting instances, the reaction of Reddit users was way faster than the one of dedicated news agencies. Essentially, each user functioned as an individual regional news relay. While based on the influence range of an event, more users participated and led to the emergence of the event. Thus, one of the highest profile functions of such an OSN analysis tool, would be to take advantage of that collective real time reporting. Which could be extended to collective problem solving, collective reaction etc. all in real time, thanks to the interaction of users with modern OSNs.

As for the social network proposal, the aim will be to study the effects of anonymity in Online Social Networks and further analyse equality and bias and promote productivity. On the one hand, it is apparent that these goals cannot be completely achieved in any form of established OSN, because a public profile creation is required and data access is restricted in all of them, on the other hand, anonymous online networks or applications that do not require a public profile, are either focused in different forms of communication, or their data collection is insufficient. Thus, we also aim to provide unrestricted data access to interested researchers and individuals with outmost respect to user privacy.

This experimental network will strive to protect user's privacy and retain an open data structure. Universal anonymity is chosen in order to try and achieve these goals and present a novel platform for OSN analysis. However, before concluding to anonymity's catalytic factor in user interaction, we should study it in a monitored environment. The proposed network will also serve as an extensive experimental platform for anonymity and/or similar concepts. From a researcher's standpoint we look forward into creating a free and open social network. The idea is constantly evolving through dialogue and

knowledge exchange. Even if the proposed network will not reach a wide user adoption, its unique character will surely provide a new research platform for computer, information and social scientists alike.

BIBLIOGRAPHY

[1] L. MD. cCallum, A review of existing needs assessment methodologies and some suggestions for change: becoming aware of ethnicity Social Networks, and Family Structure, Diss. Wayne State University, 1900.

[2] T. Ferdinand, Gemeinschaft und Gesellschaft, Ripol Classic, 1887.

[3] M. S. Feldman, "Constraints on Communication and Electronic Mail," in *ACM conference on Computer-supported cooperative work*, 1986.

[4] M. Granovetter, "The Strength of Weak Ties," *American journal of sociology,* no. 78, p. I, 1973.

[5] M. L. Markus, "Toward a “critical mass” theory of interactive media universal access, interdependence and diffusion," *Communication Research,* vol. 14, no. 5, pp. 491-511, October 1987.

[6] P. Oliver, G. Marwell and R. Teixeira, "A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action.," *American journal of Sociology*, pp. 522-556, 1985.

[7] N. S. Contractor and E. M. Eisenberg, "Communication networks and new media in organizations. Organizations and communication technology," pp. 143-172, 1990.

[8] R. E. Rice, A. E. Grant, J. Schmitz and J. Torobin, "Individual and network influences on the adoption and perceived outcomes of electronic messaging," *Social networks*, no. 12, pp. 27-55, 1990.

[9] J. Hollan and S. Stornetta, "Beyond being there," *Proceedings of the SIGCHI conference on Human factors in computing systems,* pp. 119-125, June 1992.

[10] A. Mee, "The pleasure telephone," *The Strand Magazine,* no. 44, pp. 339-345, 1898.

[11] L. E. Garton and B. Wellman, "Social impacts of electronic mail in organizations: A review of the research literature," 1993.

[12] J. D. Eveland, A. Blanchard, W. Brown and J. Mattocks, "The role of “help networks” in facilitating use of CSCW tools," in *ACM conference on Computer supported cooperative work*, 1994.

[13] J. M. Pickering and J. L. King, "Hardwiring weak ties: Interorganizational computer-mediated communication, occupational communities, and organizational change," *Organization Science*, no. 6, pp. 479-486, 1995.

[14] C. Haythornthwaite, B. Wellman and M. Mantei, "Work relationships and media use: A social network analysis," *Group Decision and Negotiation,* no. 4, pp. 193-211, 1995.

[15] H. Ogata, A. Goji, Q. Jin, Y. Yano and N. Furugori, "Distributed PeCo-Mediator: Finding Partners via Personal Connections," in *IEEE International Conference on Systems, Man, and Cybernetics*, 1996.

[16] C. Haythornthwaite, "Social network analysis: An approach and technique for the study

of information exchange," *Library & Information Science Research,* no. 4, pp. 323-342, 1996.

[17] D. Constant, L. Sproull and S. Kiesler, "The Kindness of Strangers: The Usefulness of Electronic Weak Ties for Technical Advice," *Organization Science,* vol. 7, no. 2, pp. 119 - 135, 1996.

[18] R. Kling, "Social relationships in electronic forums: Hangouts, salons, workplaces and communities," in *Computerization and controversy: Value conflicts and social choices,* 1996, pp. 426-454.

[19] B. Wellman, "An electronic group is virtually a social network," *Culture of the Internet,* vol. 4, pp. 179-205, 1997.

[20] C. Haythornthwaite and B. Wellman, "Work, friendship, and media use for information exchange in a networked organization," *Journal of the American Society for Information Science,* no. 49, pp. 1101-1114, 1998.

[21] T. Erickson, D. N. Smith, W. A. Kellogg, M. Laff, J. T. Richards and E. Bradner, "Socially translucent systems: social proxies, persistent conversation, and the design of "babble"," in *SIGCHI conference on Human Factor,* 1999.

[22] X. Fu, J. Budzik and K. J. Hammond, "Leveraging Shared Context to Facilitate Opportunistic Communication," 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.39.3648&rep=rep1&type=pdf.

[23] D. Andrews, J. Preece and M. Turoff, "A conceptual framework for demographic groups resistant to online community interaction," in *Proceedings of the 34th Annual Hawaii International Conference on System Sciences,* 2001.

[24] H. Ogata, Y. Yano, N. Furugori and Q. Jin, "Computer supported social networking for augmenting cooperation," *Computer Supported Cooperative Work,* vol. 2, no. 10, pp. 189-209, 2001.

[25] R. T. Sparrowe, R. C. Liden, S. J. Wayne and M. L. Kraimer, "Social networks and the performance of individuals and groups," *Academy of management journal,* no. 44, pp. 316-325, 2001.

[26] D. J. Watts, P. S. Dodds and M. E. Newman, "Identity and search in social networks," *Science,* no. 17, pp. 1302-1305, 2002.

[27] L. E. Garton, C. Haythornthwaite and B. Wellman, "Studying Online Social Networks," *Journal of Computer-mediated Communication,* vol. 3, no. 1, 1997.

[28] H. Ogata, T. Sueda, N. Furugori and Y. Yano, "Augmenting collaboration beyond classrooms through online social networks," in *Advanced Research in Computers and Communications in Education,* 1999.

[29] J. Howland and C. Mayer, "Tools of innovation: Supporting change through online web solutions," Honolulu, 1999.

[30] L. Kimball and H. Rheingold, "How online social networks benefit organizations," 2000.

[31] A. Q. Haase, K. Hampton, B. Wellman and J. Witte, "Does the Internet increase, decrease, or supplement social capital? Social networks, participation, and community

commitment," *American behavioral scientist,* vol. 45, pp. 436-455, 2001.

[32] S. J. Ball-Rokeach and S. Matei, "Real and virtual social ties connections in the everyday lives of seven ethnic neighborhoods," *American Behavioral Scientist,* vol. 45, pp. 550-564, 2001.

[33] J. M. Bernhardt and J. Hubley, "Health education and the Internet: the beginning of a revolution," *Health Education Research,* vol. 16, pp. 643-645, 2001.

[34] M. O'Reilly and D. Newton, "Interaction online: Above and beyond requirements of assessment," *Australian Journal of Educational Technology,* vol. 1, no. 18, pp. 57-70, 2002.

[35] S. J. Ball-Rokeach, C. Haythornthwaite, S. Matei and B. Wellman, "Belonging in geographic, ethnic, and Internet spaces," in *The Internet in everyday life*, 2002, pp. 404-430.

[36] M. R. Subramani and B. Rajagopalan, "Knowledge-sharing and influence in online social networks via viral marketing," *Communications of the ACM,* vol. 12, no. 46, pp. 300-307, 2003.

[37] M. McCaughey and M. D. Ayers, Cyberactivism: Online activism in theory and practice, Routledge, 2003.

[38] J. Suler, "The Online Disinhibition Effect," *Cyberpsychology & behavior,* vol. 7, no. 3, pp. 321-326, 2004.

[39] S. Decker and M. Frank, "The Social Semantic Desktop," DERI Technical Report, 2004.

[40] T. Hogg and L. Adamic, "Enhancing reputation mechanisms via online social networks," in *5th ACM conference on Electronic commerce*, 2004.

[41] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan and A. Tomkins, "Geographic Routing in Social Networks," in *National Academy of Sciences of the United States of America*, 2005.

[42] J. Heer and D. Boyd, "Vizster: Visualizing online social networks," in *In Information Visualization*, 2005.

[43] R. Gross and A. Acquisti, "Information Revelation and Privacy in Online Social Networks," in *ACM Workshop on Privacy in the Electronic Society*, 2005.

[44] C. Lampe, N. Ellison and C. Steinfie, "A Face (book) in the Crowd: Social Searching vs. Social Browsing," in *20th anniversary conference on Computer supported cooperative work*, Alberta, 2006.

[45] N. Ellison, C. Steinfield and C. Lam, "Spatially Bounded Online Social Networks and Social Capital," *International Communication Association,* no. 36, pp. 1-37, 2006.

[46] P. Kazienko and K. Musiał, Recommendation Framework for Online Social Networks, Berlin: Springer Berlin Heidelberg, 2006.

[47] A. Mislove, K. P. Gummadi and P. Drusc, "Exploiting Social Networks for Internet Search," in *Workshop on Hot Topics in Networks*, Irvine, 2006.

[48] A. Mislove, K. P. Gummadi, M. Marcon, P. Druschel and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM*

SIGCOMM conference on Internet Measurement, New York, 2007.

[49] Y.-Y. Ahn, S. Han, H. Kwak and S. Mo, "Analysis of topological characteristics of huge online social networking services," in *Proceedings of the 16th international conference on World Wide Web*, New York, 2007.

[50] A. N. Joinson, "Looking at, looking up or keeping up with people?: motives and use of facebook," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2008.

[51] T. A. Y. A. Y. a. S. L. C. ". 3. n. 3. (. 2.-2. Pempek, "College students' social networking experiences on Facebook," *Journal of Applied Developmental Psychology,* vol. 30, no. 3, pp. 227-238, 2009.

[52] B. Viswanath, A. Mislove, M. Cha and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proceedings of the 2nd ACM workshop on Online social networks*, 2009.

[53] N. B. C. S. a. C. L. Ellison, "Connection strategies: Social capital implications of Facebook-enabled communication practices," *New Media & Society,* vol. 14, no. 3, pp. 1-20, 2011.

[54] B. A. Huberman, D. M. Romero and F. Wu, "Social networks that matter: Twitter under the microscope," 2008. [Online]. Available: http://arxiv.org/abs/0812.1045.

[55] B. J. Jansen, M. Zhang, K. Sobel and A. Chowdury, "Twitter power: Tweets as electronic word of mouth."," *Journal of the American society for information science and technology*, vol. 60, no. 11, pp. 2169-2188, 2009.

[56] H. Kwak, C. Lee, H. Park and S. Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, 2010.

[57] J. Bollen, H. Mao and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science,* vol. 2, no. 1, pp. 1-8.

[58] J. Raacke and J. Bonds-Raacke, "MySpace and Facebook: Applying the uses and gratifications theory to exploring friend-networking sites," *Cyberpsychology & behavior*, vol. 11, no. 2, pp. 169-174, 2008.

[59] S. Hinduja and J. W. Patchin, "Personal information of adolescents on the Internet: A quantitative content analysis of MySpace."," *Journal of Adolescence,* vol. 31, no. 1, pp. 125-146, 2008.

[60] U. Pfeil, R. Arjan and P. Zaphiris, "Age differences in online social networking–A study of user profiles and the social capital divide among teenagers and older users in MySpace," *Computers in Human Behavior,* vol. 6, no. 1, pp. 643-654, 2009.

[61] C. L. Kujath, "Facebook and MySpace: Complement or substitute for face-to-face interaction?," *Cyberpsychology, Behavior, and Social Networking*, vol. 14, no. 1-2, pp. 75-78, 2011.

[62] F. Benevenuto, T. Rodrigues, M. Cha and V. Almeida, "Characterizing user navigation and interactions in online social networks," *Information Sciences*, no. 195, pp. 1-24, 2012.

[63] X. Yang, H. Steck and Y. Liu, "Circle-based recommendation in online social networks." In," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

[64] G. Veletsianos and C. Navarrete, "Online social networks as formal learning environments: Learner experiences and activities," *The International Review of Research in Open and Distance Learning*, vol. 13, no. 1, pp. 144-166, 2012.

[65] S. Jahid, S. Nilizadeh, P. Mittal, N. Borisov and A. Kapadia, "DECENT: A decentralized architecture for enforcing privacy in online social networks," in *Pervasive Computing and Communications Workshops*, 2012.

[66] M. Madejski, M. Johnson and S. M. Bellovin, "A study of privacy settings errors in an online social network," in *Pervasive Computing and Communications Workshops*, 2012.

[67] H. Hu, G.-J. Ahn and J. Jorgensen, "Multiparty access control for online social networks: model and mechanisms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1614-1627, 2013.

[68] F. Heart, A. McKenzie, J. M. McQuillan and D. C. Walden, ARPANET Completion Report, BBN, 1978.

[69] T. O'Reilly, "What is Web 2.0: Design patterns and business models for the next generation of software," *Communications & Strategies*, no. 1, p. 17, 2007.

[70] N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210-230, 2007.

[71] K. Colstad and E. Lipkin, "Community memory: a public information network," *SIGCAS Computers and Society*, vol. 6, no. 4, pp. 6-7, December 1975.

[72] L. Euler, "Solutio problematis ad geometriam situs pertinentis," *Commentarii academiae scientiarum Petropolitanae* , pp. 128-140, 1741.

[73] M. E. Newman, "Mixing patterns in networks," *Physical Review E*, no. 2, 2003.

[74] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129-1164, 1991.

[75] M. P. Stumpf, C. Wiuf and R. M. May, "Subnets of scale-free networks are not scale-free: sampling properties of networks," in *Proceedings of the National Academy of Sciences of the United States of America*, 2005.

[76] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[77] J. Leskovec, J. Kleinberg and C. Faloutsos, "Graphs over time: Densification laws, shrinking diamaters and possible explanations," in *Proceeding of SIGKDD international conference on Knowledge discovery and data mining*, 2005.

[78] S. H. Lee, P.-J. Kim and H. Jeong, "Statistical properties of sampled networks," *Physical Review E*, vol. 73, no. 1, 2006.

[79] V. Pareto, Course of Political Economy, 1896.

[80] L. A. Adamic, R. M. Lukose, A. R. Puniyani and B. A. Huberman, "Search in power-law

networks," *Physical review E,* vol. 64, no. 4, p. 046135, 2001.

[81] B. A. Huberman and L. A. Adamic, "Internet: growth dynamics of the world-wide web," *Nature 401,* pp. 131-131, 1999.

[82] C. Castellano, S. Fortunato and V. Loreto, "Statistical physics of social dynamics," *Reviews of modern physics,* vol. 81, no. 2, pp. 591-646, 2009.

[83] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science,* vol. 286, no. 5439 , pp. 509-512, 1999.

[84] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, "Graph structure in the web," *Computer networks,* vol. 33, no. 1, pp. 309-320, 2000.

[85] M. E. Newman, D. J. Watts and S. H. Strogatz, "Random graph models of social networks," in *Proceedings of the National Academy of Sciences,* 2002.

[86] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J.-H. Cui and A. G. Percus, "Reducing large internet topologies for faster simulations," *NETWORKING 2005. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems,* pp. 328-341, 2005.

[87] J. Leskovec, K. J. Lang, A. Dasgupta and M. W. Mahoney, "Statistical properties of community structure in large social and information networks," in *Proceedings of the 17th international conference on World Wide Web,* ACM, 2008.

[88] F. R. Chung, Spectral graph theory, vol. 92, American Mathematical Society, 1997.

[89] R. Zou and L. B. Holder, "Frequent subgraph mining on a single large graph using sampling techniques," in *Proceedings of the eighth workshop on mining and learning with graphs,* 2010.

[90] N. Kashtan, S. Itzkovitz, R. Milo and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics,* vol. 20, no. 11, pp. 1746-1758, 2004.

[91] P. Hintsanen and H. Toivonen, "Finding reliable subgraphs from large probabilistic graphs," *Data Mining and Knowledge Discovery,* vol. 17, no. 1, pp. 3-23, 2008.

[92] B. Krishnamurthy, P. Gill and M. Arlitt, "A few chirps about twitter," in *Proceedings of the first workshop on Online social networks,* 2008.

[93] A. H. Rasti, M. Torkjazi, R. Rejaie and D. Stutzbach, *Evaluating sampling techniques for large dynamic graphs,* vol. 1, University of Oregon, 2008, pp. 1-14.

[94] M. Gjoka, M. Kurant, C. T. Butts and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of osns," in *Proceedings of 29th Conference on Information communications,* 2010.

[95] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement,* 2010.

[96] M. Gjoka, C. T. Butts, M. Kurant and A. Markopoulou, "Multigraph sampling of online

social networks," *Selected Areas in Communications,* vol. 29, no. 9, pp. 1893-1905, 2011.

[97] P. Noordhuis, M. Heijkoop and A. Lazovik, "Mining twitter in the cloud: A case study," in *Proceedings of the IEEE 3rd International Conference in Cloud Computing*, 2010.

[98] D. H. Chau, S. Pandit, S. Wang and C. Faloutsos, "Parallel crawling for online social networks," in *In Proceedings of the 16th international conference on World Wide Web*, 2007.

[99] S. Ye, J. Lang and F. Wu, "Crawling online social graphs," in *Proccedings of 12th International Asia-Pacific on Web Conference*, 2010.

[100] T. Feder and R. Motwani, "Clique partitions, graph compression and speeding-up algorithms," *Journal of Computer and System Sciences,* vol. 51, no. 2, pp. 261-272, 1995.

[101] M. Adler and M. Mitzenmacher, "Towards compressing web graphs," in *Proceedings of Data Compression Conference*, 2001.

[102] L. Katzir, E. Liberty, O. Somekh and I. A. Cosma, "Estimating sizes of social networks via biased sampling," *Internet Mathematics*, 2014.

[103] M. Kurant, C. T. Butts and A. Markopoulou, "Graph Size Estimation," arXiv, 2012.

[104] A. Clauset, C. R. Shalizi and M. E. Newman, "Power-law distributions in empirical data," *SIAM review,* vol. 51, no. 4, pp. 661-703, 2009.

[105] A. Klaus, S. Yu and D. Plenz, "Statistical analyses support power law distributions found in neuronal avalanches."," *PloS one,* vol. 6, no. 5, 2011.

[106] R. Ihaka and R. Gentlema, "R: a language for data analysis and graphics," *Journal of computational and graphical statistics,* vol. 5, no. 3, pp. 299-314, 1996.

[107] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Systems,* vol. 1695, no. 5, 2006.

[108] B. Berelson and H. Gaudet, The people’s choice: How the voter makes up his mind in a presidential campaign, 1944.

[109] E. Katz and P. F. Lazarsfeld, Personal Influence, The part played by people in the flow of mass communications, Transaction Publishers, 1970.

[110] E. Bakshy, I. Rosenn, C. Marlow and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*, 2012.

[111] A. B. Rajyalakshmi, S. Das and R. M. Tripathy, "Topic Diffusion and Emergence of Virality in Social Networks," 2012.

[112] S. Jurvetson and T. Draper, "What is viral marketing?," 1 January 1997. [Online]. Available: http://dfj.com/news/article_26.shtml.

[113] P. Lance and G. G. J, "From subservient chickens to brawny men: A comparison of viral advertising to television advertising," *Journal of Interactive Advertising,* vol. 2, pp. 4-33, 2006.

[114] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen and N. Glance, "Cost-

effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007.

[115] D. T. Allsop, B. R. Bassett and J. A. Hoskins, "Word-of-mouth research: principles and applications," *Journal of Advertising Research*, vol. 4, p. 398, 2007.

[116] J. Berger and K. Milkman, *Social transmission, emotion, and the virality of online content*, Wharton Research Paper, 2010.

[117] M. Gomez Rodriguez, J. Leskovec and A. Krause, "Inferring networks of diffusion and influence," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010.

[118] J. L. Iribarren and E. Moro, "Affinity Paths and information diffusion in social networks," *Social networks*, vol. 2, pp. 134-142, 2011.

[119] A. Goyal, F. Bonchi and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010.

[120] D. Ienco, F. Bonchi and C. Castillo, "The Meme Ranking Problem: Maximizing Microblogging Virality," in *IEEE International Conference on Data Mining Workshops*, 2010.

[121] J. Yang and S. Counts, "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter," in *Proceedings of the International Conference on Weblogs and Social Media*, 2010.

[122] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," in *Proceedings of the 19th international conference on World wide web*, 2010.

[123] C. Tucker, "Ad virality and ad persuasiveness," SSRN, 2014.

[124] K. Nahon, J. Hemsley, S. Walker and M. Hussain, "Blogs: spinning a web of virality," in *Proceedings of the 2011 iConference*, 2011.

[125] M. Guerini, C. Strapparava and G. Özbal, "Exploring Text Virality in Social Networks," in *Proceedings of International Conference on Weblogs and Social Media* , 2011.

[126] L. K. Hansen, A. Arvidsson, F. Å. Nielsen, E. Colleoni and M. Etter, "Good friends, bad news-affect and virality in twitter," *Future information technology*, pp. 34-43, 2011.

[127] F. Bonchi, "Influence Propagation in Social Networks: A Data Mining Perspective," *IEEE Intelligent Informatics Bulletin*, vol. 1, pp. 8-16, 2011.

[128] D. M. Romero, W. Galuba, S. Asur and B. A. Huberman, "Influence and passivity in social media," *Machine learning and knowledge discovery in databases*, pp. 18-33, 2011.

[129] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis and A. Ukkonen, "Sparsification of influence networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.

[130] D. M. Romero, B. Meeder and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion

on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011.

[131] J. Yang and J. Leskovec, "Patterns of Temporal Variation in Online Media," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011.

[132] M. Guerini, A. Pepe and B. Lepri, "Do Linguistic Style and Readability of Scientific Abstracts Affect their Virality?," in *Proceedings of International Conference on Weblogs and Social Media*, 2012.

[133] Y. Li, S. Huang, C. Fan and G. Yang, "The Selection of Information Diffusion Monitoring Nodes in Directed Online Social Networks," in *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*, 2013.

[134] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in *Proceedings of the 21st international conference companion on World Wide Web*, 2012.

[135] A. Karnik, A. Saroop and V. Borkar, "On the diffusion of messages in on-line social networks," *Performance Evaluation*, vol. 4, pp. 271-285, 2013.

[136] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer and A. Flammini, "The role of information diffusion in the evolution of social networks," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.

[137] A. Portes and J. Sensenbrenner, "Embeddedness and immigration: Notes on the social determinants of economic action," *American Journal of Sociology*, pp. 1320-1350, 1993.

[138] M. Girvan and M. E. Newman, "Community structure in social and biological networks." 99, no. 12 (): ," *Proceedings of the National Academy of Sciences*, vol. 12, pp. 7821-7826, 2002.

[139] B. Wellman and B. Leighton, "Networks, Neighborhoods, and Communities Approaches to the Study of the Community Question." 14, no. 3 ():," *Urban Affairs Review*, vol. 3, no. 14, pp. 363-390, 1979.

[140] N. Long, Development Sociology: Actor Perspectives, Routledge, 2001.

[141] M. Jacomy, S. Heymann, T. Venturini and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization," *Medialab center of research*, p. 560, 2011.

[142] T. Parsons, The social structure of the family, 1949.

[143] F. I. Nye, Family relationships and delinquent behaviour, 1958.

[144] H. Menzel and E. Katz, "Social relations and innovation in the medical profession: the epidemiology of a new drug."," *Public Opinion Quarterly*, vol. 4, no. 19, pp. 337-352, 1955.

[145] J. A. Barnes, Class and committees in a Norwegian island parish, Plenum, 1954.

[146] R. V. Speck, "Psychotherapy of the social network of a schizophrenic family," *Family process*, vol. 2, no. 6, pp. 208-214, 1967.

[147] A. R. Radcliffe-Brown, " On social structure," *Journal of the Anthropological Institute of Great Britain and Ireland,* pp. 1-12, 1940.

[148] M. L. Cadwallader, "The cybernetic analysis of change in complex social organizations," *American Journal of Sociology,* pp. 154-157, 1959.

[149] J. L. Moreno, "Foundations of sociometry: An introduction," *Sociometry ,* pp. 15-35, 1941.

[150] B. H. Raven, ""The bases of power: Origins and recent developments," *Journal of social issues,* vol. 4, no. 49, pp. 227-251, 1993.

[151] A. W. Gouldner, "The norm of reciprocity: A preliminary statement," *American sociological review,* pp. 161-178, 1960.

[152] B. N. Adams, "Interaction theory and the social network," *Sociometry ,* pp. 64-78, 1967.

[153] P. Craven and B. Wellman, "The Network City," *Sociological inquiry,* Vols. 3-4, no. 43, pp. 57-88, 1973.

[154] B. Wellman, "Structural analysis: From method and metaphor to theory and substance," *Contemporary Studies in Sociology,* no. 15, pp. 19-61, 1997.

[155] P. D. Killworth and H. R. Bernard, "Informant accuracy in social network data," *Human Organization,* vol. 3, pp. 269-286, 1976.

[156] J. Fulk, C. W. Steinfield, J. Schmitz and J. G. Power, "A social information processing model of media use in organizations," *Communication research,* vol. 5, pp. 529-552, 1987.

[157] A. Clement, "Cooperative support for computer work: a social perspective on the empowering of end users," in *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, 1990.

[158] B. Shneiderman, Designing the user interface-strategies for effective human-computer interaction, Pearson Education India, 1986.

[159] B. Wellman and S. Wortley, "Different strokes from different folks: Community ties and social support," *American journal of Sociology,* pp. 558-588, 1990.

[160] B. Wellman, "Which types of ties and networks provide what kinds of social support," *Advances in group processes,* no. 9, pp. 207-235, 1992.

[161] N. Hardy and E. D. Tribble, "The digital silk road," *In Tech. Rep. Agorics, Inc,* 1993.

[162] W. E. Mackay, ""Ethics, lies and videotape…." In," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 138-145, 1995.

[163] M. E. Burkhardt and D. J. Brass, "Changing patterns or patterns of change: The effects of a change in technology on social network structure and power," *Administrative science quarterly,* pp. 104-127, 1990.

[164] T. Postmes, R. Spears, K. Sakhel and D. D. Groot, "Social influence in computer-mediated communication: The effects of anonymity on group behavior," *Personality and Social Psychology Bulletin ,* vol. 10, no. 27, pp. 1243-1254, 2001.

[165] E. O. Laumann, Perspectives on social network research, 1979, pp. 379-402.

[166] P. J. Carrington, J. Scott and S. Wasserman, Models and methods in social network

analysis, vol. 28, Cambridge university press, 2005.

[167] S. E. Kiesler, Culture of the Internet, 1997.

[168] C. Dwyer, S. Hiltz and K. Passerini, "Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace," in *2007,* AMCIS 2007 Proceedings.

[169] A. Acquisti and R. Gross, "Imagined communities: Awareness, information sharing, and privacy on the Facebook," *In Privacy enhancing technologies,* pp. 36-58, 2006.

[170] D. Boyd, "Facebook's Privacy Trainwreck," *Convergence: The International Journal of Research into New Media Technologies,* vol. 14, no. 1, pp. 13-20, 2008.

[171] Y. Liu, K. P. Gummadi, B. Krishnamurthy and A. Mislove, "Analyzing facebook privacy settings: user expectations vs. reality," in *In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference,* 2011.

[172] H. R. Lipford, A. Besmer and J. Watson, "Understanding Privacy Settings in Facebook with an Audience View," *Usability, Psychology, and Security,* vol. 8, pp. 1-8, 2008.

[173] H. Mao, X. Shuai and A. Kapadia, "Loose tweets: an analysis of privacy leaks on twitter," *In Proceedings of the 10th annual ACM workshop on Privacy in the electronic society,* pp. 1-12, 2011.

[174] L. Humphreys, P. Gill and B. Krishnamurthy, "How much is too much? Privacy issues on Twitter," *In Conference of International Communication Association,* 2010.

[175] E. De Cristofaro, C. Soriente, G. Tsudik and A. Williams, "Hummingbird: Privacy at the time of twitter," *In IEEE Symposium of Security and Privacy,* pp. 285-299, 2012.

[176] M. Chau and J. Xu, "Mining communities and their relationships in blogs: A study of online hate groups," *International Journal of Human-Computer Studies,* vol. 1, no. 65, pp. 57-70, 2007.

[177] J. N. Cummings, B. Butler and R. Kraut, "The Quality of Online Social Relationships," *Communications of the ACM,* vol. 7 , no. 45, pp. 103-108, 2002.

[178] N. B. Ellison, C. Steinfield and C. Lampe, "The benefits of Facebook “friends:” Social capital and college students’ use of online social network sites," *Journal of Computer-Mediated Communication,* vol. 4, no. 12, pp. 1143-1168, 2007.

[179] T. Iba, K. Nemoto, B. Peters and P. A. Gloor, "Analyzing the creative editing behavior of Wikipedia editors: Through dynamic social network analysis," *Procedia-Social and Behavioral Sciences,* vol. 4, no. 2, pp. 6441-6456, 2010.

[180] J. M. Darley and B. Latane, "Bystander intervention in emergencies: diffusion of responsibility," *Journal of personality and social psychology,* vol. 4 Part I, no. 8, pp. 377-383, 1968.

[181] M. A. Wallach, N. Kogan and D. J. Bem, "Diffusion of responsibility and level of risk taking in groups," *The Journal of Abnormal and Social Psychology,* vol. 3, no. 68, pp. 263-274, 1964.

[182] M. Faloutsos, P. Faloutsos and C. Faloutsos, "On power-law relationships of the internet topology," *In ACM SIGCOMM Computer Communication Review,* vol. 29, no. 4, pp.

251-262, 1999.

[183] L. A. Adamic and B. A. Huberman, "Power-law distribution of the world wide web," *Science 287*, no. 5461, pp. 2115-2115, 2000.

[184] S. L. Johnson, S. Faraj and S. Kudaravalli, "Emergence of power laws in online communities: the role of social mechanisms and preferential attachment," *Mis Quarterly*, vol. 3, no. 38, pp. 795-808, 2014.

[185] A. Sala, H. Zheng, B. Y. Zhao, S. Gaito and G. P. Rossi, "Brief Announcement: Revisiting the Power-law Degree," 2010.

[186] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101-113, 2004.

[187] P. M. DeMarzo, J. Zwiebel and D. Vayanos, *Persuasion bias, social influence, and uni-dimensional opinions*, MIT Sloan Working Paper 4339-01, 2001, pp. 4339-01.

[188] Y.-R. Lin, J. P. Bagrow and D. Lazer, "More voices than ever? quantifying media bias in networks," arXiv preprint, 2011.

[189] S. Counts, "Attention and Bias in Social Information Networks," 2011.

[190] E. Kiciman, "Microsoft Research," [Online]. Available: http://research.microsoft.com/en-us/um/cambridge/events/mls2013/downloads/investigating_bias.pdf.

[191] K. L. Nowak and C. Rauh, "The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction," *Journal of Computer-Mediated Communication*, vol. 11, no. 1, pp. 153-178, 2005.

[192] D. Westerman, P. R. Spence and B. V. D. Heide, "A social network as information: The effect of system generated reports of connectedness on credibility on Twitter," *Computers in Human Behavior*, vol. 28, no. 1, pp. 199-206, 2012.

[193] M. J. Metzger, A. J. Flanagin and R. B. Medders, "Social and heuristic approaches to credibility evaluation online," *Journal of Communication*, vol. 60, no. 3, pp. 413-439, 2010.

[194] J. K. Ayeh, N. Au and R. Law, ""Do we believe in TripAdvisor?" Examining credibility perceptions and online travelers' attitude toward using user-generated content," *Journal of Travel Research*, p. 0047287512475217, 2013.

[195] A. Gupta, P. Kumaraguru, C. Castillo and P. Meier, "TweetCred: A Real-time Web-based System for Assessing Credibility of Content on Twitter," arXiv, 2014.

[196] M. Schmierbach and A. Oeldorf-Hirsch, "A little bird told me, so I didn't believe it: Twitter, credibility, and issue perceptions," *Communication Quarterly*, vol. 60, no. 3, pp. 317-337, 2012.

[197] S.-a. a. B. J. M. Knight, "Developing a framework for assessing information quality on the World Wide Web," *Informing Science: International Journal of an Emerging Transdiscipline*, vol. 8, no. 5, pp. 159-172, 2005.

[198] B. F. Loscio, M. Batista, D. Souza and A. C. Salgado, "Using information quality for the identification of relevant web data sources: a proposal," in *14th International*

Conference on Information Integration and Web-based Applications & Services, 2012.

[199] M. Haklay, "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets," *Environment and planning. B, Planning & design*, vol. 37, no. 4, p. 682, 2010.

[200] E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. a. H. C. Cagnina, B. Stein and M. Granitzer, "Measuring the quality of web content using factual information," in *2nd joint WICOW/AIRWeb workshop on web quality*, 2012.

[201] C. Cappiello, "On the Role of Data Quality in Improving Web Information Value."," in *24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee*, 2015.

[202] M. Kc, M. Hagenbuchner and A. C. Tsoi, "Quality information retrieval for the World Wide Web," in *International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.

[203] F. Bodendorf and C. Kaiser, "Detecting opinion leaders and trends in online social networks," in *2nd ACM workshop on Social web search and mining*, 2009.

[204] J. W. Patchin and S. Hinduja, "Trends in online social networking: Adolescent use of MySpace over time," *New Media & Society*, vol. 12, no. 2, pp. 197-216, 2010.

[205] S. Wright, "Government-run Online Discussion Fora: Moderation, Censorship and the Shadow of Control," *The british journal of politics & international relations*, vol. 8, no. 4, pp. 550-568, 2006.

[206] C. S. Andreassen, T. Torsheim, G. S. Brunborg and S. Pallesen, "Development of a facebook addiction scale"," *Psychological reports*, vol. 110, no. 2, pp. 501-517, 2012.

[207] D. Karaiskos, E. Tzavellas, G. Balta and T. Paparrigopoulos, "P02-232-Social network addiction: a new clinical disorder?," *European Psychiatry*, vol. 25, p. 855, 2010.

[208] M. D. Griffiths, "Facebook Addiction: Concerns, Criticism, and Recommendations - a Response To Andreassen and Colleagues," *Psychological Reports*, vol. 110, no. 2, pp. 518-520, 2012.