Data mining of social networks represented as graphs

David F. Nettleton[a,b]

[a]*Universitat Pompeu Fabra, Barcelona, Spain*

[b]*IIIA-CSIC, Bellaterra, Spain*

Abstract

In this survey we review the literature and concepts of the data mining of social networks, with special emphasis on their representation as a graph structure. The survey is divided into two principal parts: firstly we conduct a survey of the literature which forms the 'basis' and background for the field; and secondly we define a set of 'hot topics' which are currently in vogue in congresses and the literature. The 'basis' or background part is divided into four major themes: graph theory, social networks, online social networks and graph mining. The graph mining theme is organized into ten subthemes. The second, 'hot topic' part, is divided into five major themes: communities, influence and recommendation, models metrics and dynamics, behaviour and relationships, and information diffusion.

1. Introduction

The analysis of social networks has recently experienced a surge of interest by researchers, due to different factors, such as the popularity of online social networks (OSNs), their representation and analysis as graphs, the availability of large volumes of OSN log data, and commercial/marketing interests. OSNs also present interesting research challenges such as the search/matching of similar sub-graphs, community analysis/modelling, user classification and information propagation. Hence, OSN data analysis has a great potential for researchers in a diversity of disciplines. However, we propose that OSN analysis should be placed in the context of its sociological origins and its basis in graph theory. Thus, we have devised a survey which firstly presents the key historical and base research ideas and the different associated themes, and secondly presents a selection of the latest research and tendencies taken from international conferences.

Graph Mining of on-line social networks is a relatively new area of research which however has a solid base in classic graph theory, computational cost considerations, and sociological concepts such how individuals interrelate, group together and follow one another.

For the purposes of the survey, we will divide the base themes as follows: graph theory, social networks, OSNs and SN dataset analysis, and graph mining. The 'graph mining' theme is divided into sub-themes as is shown in Fig. 1. Then the 'hot topics' are divided into five sub-themes, as is illustrated in Fig. 1. The 'hot topic' themes were selected by classifying the papers found in recent editions of four major conferences: WWW 2012, ICSWM 2012, WOSN 2010 and WCCI 2012.

The structure of the paper is as follows: Section Two consists of a survey of the four major 'base' themes and ten sub-themes, highlighting the key concepts and authors. Then in Section Three we present five 'hot topics' in which we summarize a selection of the latest research. Section Four concludes with a summary of the survey and of the identified key tendencies. In Fig. 1 we see a schematic representation of the structure of the complete survey.

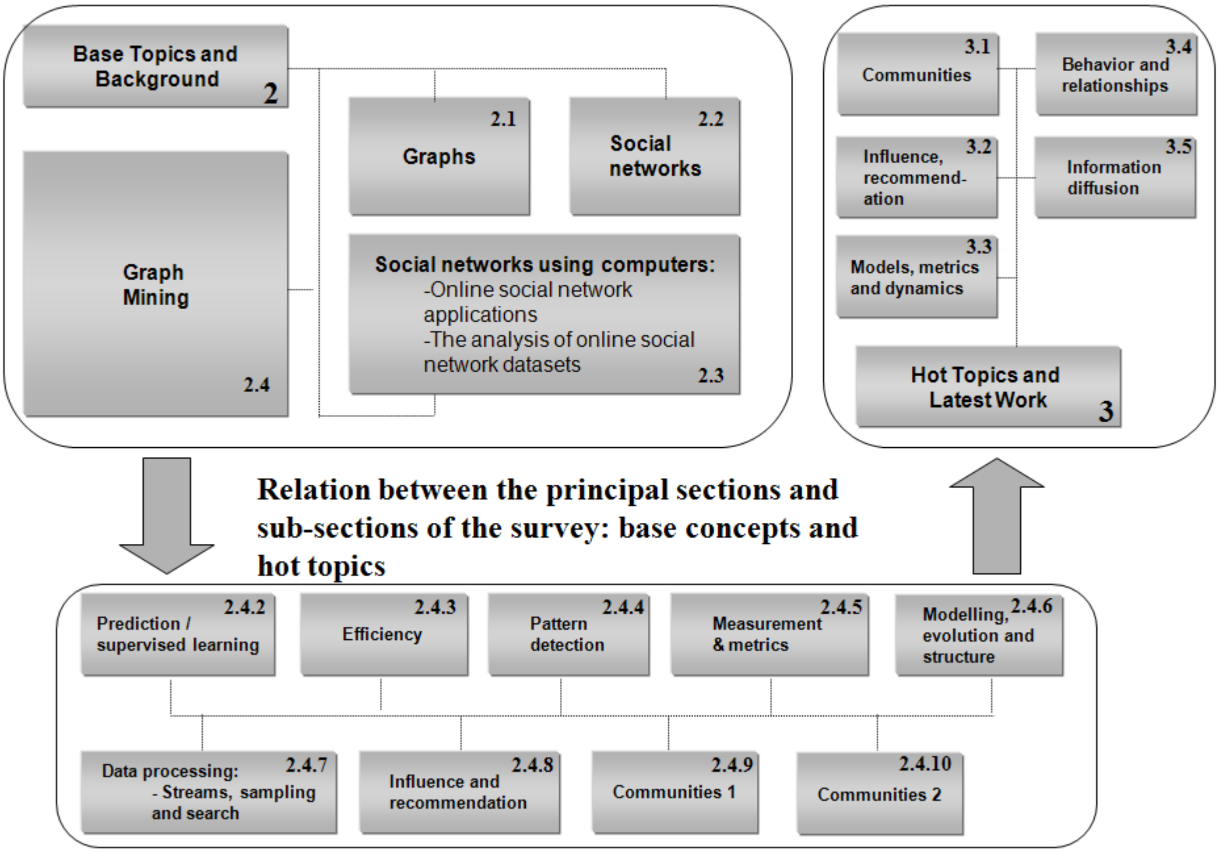

Base Topics and Background 2
Graphs 2.1
Social networks 2.2
Social networks using computers:
-Online social network applications
-The analysis of online social network datasets 2.3
Graph Mining 2.4
Communities 3.1
Influence, recommend-ation 3.2
Models, metrics and dynamics 3.3
Behavior and relationships 3.4
Information diffusion 3.5
Hot Topics and Latest Work 3
Relation between the principal sections and sub-sections of the survey: base concepts and hot topics
Prediction / supervised learning 2.4.2
Efficiency 2.4.3
Pattern detection 2.4.4
Measurement & metrics 2.4.5
Modelling, evolution and structure 2.4.6
Data processing: - Streams, sampling and search 2.4.7
Influence and recommendation 2.4.8
Communities 1 2.4.9
Communities 2 2.4.10

Fig. 1. Scope of the Survey indicating the division of base topics and hot topics

2. Base survey

In this Section we consider the base themes related to graph mining of OSNs: graph theory, social networks, online social networks and graph mining.

2.1. Graphs

In this section we will summarize some of the key abstract concepts of graphs. We will see that graph mining has a solid basis in classical graph theory.

In general, a graph G is represented as G(V, E) where V is a set of vertices (or nodes) and E is a set of edges (or links) connecting some vertex pairs in V. Statistically, a graph can be characterized by derived values such as the average degree of the nodes and the average path length between nodes. Additional characteristics are the graphs diameter, the number of triangles, the number of isomorphisms and the clustering coefficient, among others.

In Fig. 2 we see an elementary graph with five vertices and five edges. As there are no arrows, we assume it is undirected, and as the edges have no additional information attached we assume it is un-weighted. We see that nodes A, B and D have degree 2, node C has degree 3 and node E has degree 1, hence the degree sequence is {1, 2, 2, 2, 3}.

In this survey we are more interested in a graph as an abstract data type rather than a mathematical entity, the former being used to represent the latter. Different algorithms exist which perform higher level operations on graphs, such as finding its degree, finding the connectivity between its neighbours (clustering coefficient), finding a path between two nodes, (using depth-first search or breadth-first search), or finding the shortest path from one node to another. Refer to [1] for a general introduction to different types of graph algorithms which are relevant to OSNs.

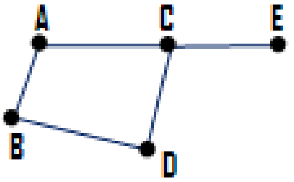


Fig. 2. Simple graph with five vertices and five edges.

A list of typical lower level graph processing operations could be the following: 'adjacent', tests if the exists an edge between two nodes; 'neighbours', finds all the nodes which have an edge with a given node; 'add', adds an edge between two nodes; 'delete', deletes an edge between two nodes; 'get' and 'set' values associated with nodes: 'get' and 'set' values associated with edges.

How to represent a graph in computer memory is a key issue, due to the potentially high computational cost of many of the higher level operators we wish to perform. Two of the most popular data structures are 'adjacency lists' and 'adjacency matrices'. Refer to [2] for more details about these structures.

Two other data structures, 'incidence lists' and 'incidence matrices' are similar to the former, with the distinction that the information stored indicates if edges and vertices are incident.

With respect to computational cost, an adjacency list is preferred when the graph connectivity is sparse, whereas an adjacency matrix is preferred if the graph is dense[2].

There are many types of graphs: directed (digraphs), undirected, graphs with weights on the edges, vertices or both, random, bipartite, and so on. In the current survey, we will principally consider non-directed graphs, directed graphs and some graphs with edge weights. An 'undirected graph' has no information about the direction or flow between nodes. That is, the edge between two vertices A and B is identical to the edge between vertices B and A. A 'directed graph', on the other hand, *does* include directional information. Each edge will have a direction associated with it, which can be unidirectional A→B or bidirectional A↔B. A 'weighted graph' includes additional information associated with an edge or a vertex. The meaning of the weight depends on the data domain, such as the number of telephone calls between two people in the last month, or the cost, benefit, capacity, and so on.

Isomorphism: A key property of interest in graphs is the isomorphic property, which refers to an exact match between two given graphs, in terms of structure, dimensionality, connectivity and mapping of corresponding nodes. Two graphs $G_1 = (V, E_1)$, $G_2 = (V, E_2)$ are designated as being isomorphic if a permutation p exists such that $p(G_1) = G_2$. That is, with the same set of vertices, the edges of G_1 can be rearranged to fit G_2.

In Fig. 3 we see an example of two graphs, one of which is an isomorphism of the other, by the permutation {(A, V), (B, W), (C, X), (D, Z), (E, Y)}.

Graph matching is a key activity in different pattern recognition applications, in which high and low level information may be represented. However, graph matching has a high computational cost, and the task of finding isomorphic sub-graphs is a problem which is NP-complete. Later, in Section 2.4.8.3 we will look at some specific graph matching and search algorithms.

For those readers who wish to enter into more detail about graph theory, West in [3] reviews the following topics: trees and distance, matching and factors, connectivity and paths, graph colouring, edges and cycles and planar graphs. An introduction to graphs and networks from a more theoretical viewpoint is found in [4], which serves as a useful and simply described reference for common graph metrics and topologies. Definitions are given for directed and undirected graphs, unconnected graphs and connected components, complete and star graphs and lattices. The following metrics are considered: clustering coefficient, average path length, centrality, and a degree distribution function. Two theoretical models are evaluated: (i) a lattice structure with shortcuts, and (ii) incremental graph evolution.

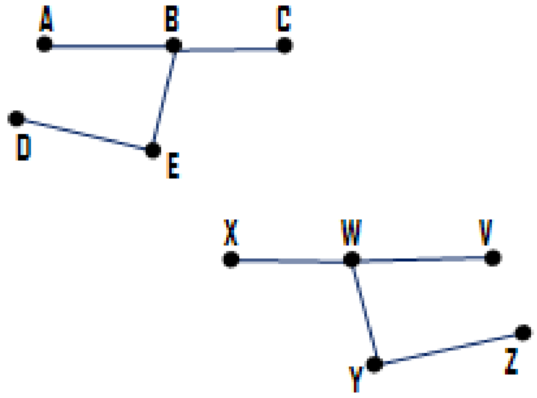


Fig. 3. Graph Isomorphisms: the upper and lower graphs are isomorphic if an adequate mapping can be defined between the two.

2.2. Social networks

A social network is a social structure comprised of a set of participants (individuals, organizations, …) and the mutual ties between these participants. The vision of a social structure as a network facilitates the understanding and analysis of this structure, such as the identification of local and global characteristics, influential participants and the dynamics of networks [5]. Social network analysis is an interdisciplinary endeavour including such fields as social psychology, sociology, statistics, and graph theory.

Georg Simmel [6][7] was one of the earliest to define structural theories in sociology, such as the dynamics of triads (subgroups involving three participants) and the development of individualism. As an example of triad dynamics consider three persons A, B, and C, in which person A has a direct dyadic relation with C and an indirect relation with C via B. In this case, person B may act to influence the relation between persons A and C.

Commencing in the 1930's, Jacob L. Moreno developed a representation called 'sociograms' which facilitated the study of interpersonal relationships, choices and preferences within groups [8]. A 'sociogram' is a diagram of the structure and patterns of group interactions, which can be based on criteria such as social relations, channels of influence, lines of communication, and so on.

During the 1950s a mathematical formalization was developed for social networks which became the basis for modern social and behavioural sciences, as well as complex network analysis in general [5].

Mc Pherson, in [9] combines a sociological focus to social network analysis with mathematical and graph theory approaches. McPherson considers the hypothesis that social network structures are influenced by the idea that similar persons are attracted to each other (the colloquial term "birds of a feather" is mentioned). Demographic factors are considered, such as race and ethnicity, gender, age, religion, education, occupation and social class. For organizations/institutions with a predefined structure, the individuals' position/role in the network is said to be a key aspect, especially if the organization is hierarchical. With reference to behavioural characteristics, these are especially important for 'affinity grouping', for example, in the case of teenagers where achievement or delinquency can create affinities, which are socially positive or negative, respectively. More ephemeral factors are defined as being attitudes, abilities, beliefs and aspirations.

In terms of the causes of "homophily" (i.e. love of the same), geography is stated as being a key factor, together with family ties, and organizational foci (school, work, voluntary organization participation) for non-kin ties. The simple conclusion is that in general people tend to associate with others who are similar to them. Finally, tie-dissolution is considered, that is, the causes of why people break ties. It is stated that weaker ties, such as those characterized by cross-gender, cross-race, or a wide age difference, are more likely to dissolve. In summary, the strongest sociological grouping factors are found to be gender, age, religion and education, with secondary factors being occupation, network position, behaviour patterns and intrapersonal values.

A research work which combines a sociological viewpoint with graph theory concepts of structures and graph generator models and their parameterization, is that of Robins et al. [10]. The central theme of this wide ranging paper, is that the overall structure of a large graph is determined by its structure at a ***local level***, specifically in terms of the proportion of a given number of predefined structures. The authors evaluate different graph model generators, in terms of the frequency of occurrence of these structures in the whole generated graph. One interesting observation is that, during graph evolution, a phase transition occurs at a given "temperature" at which regular structure formation gives way to a stochastic (random, probabilistic, natural) behaviour.

Robins et al. cite a specific analogy of **local behaviour** and **global structure** in a real social network, that of the **Medici family of 15th century Florence**. It is proposed that this family became a key player in the general social network of Florence, because they were at the centre of a star-like structure of marriage and business alliances. Their efficient networking capability with the other key players in the network was simpler and faster than the complex relationships which existed between other key players, who were their political rivals, and were also mutual rivals.

However, it is stated that this success on a global level was due to outcomes of local social processes, such as marriage ties and business partnerships. Thus, the Medici did not need a global vision of the network in order to eventually dominate it. In graph theory terms, an overall structural balance was achieved by local triadic structures. Also, the authors propose that the small world phenomenon is a global property of the graph which is a consequence of repeated local structures, such as the prevalence of short paths, thus giving a low overall average path length.

In order to model these observations Robins et al. proposed different graph generators, such as one based on random edge addition and triangle formation, which promotes a clustering tendency. Another mechanism for "growing networks" is that of "preferential attachment", in which a new node is attached to existing nodes with a probability which depends on the distribution of the degrees of the existing nodes. However, in order for this method to work optimally, it is necessary to know the distribution across the whole network. If this information cannot be obtained, then it can be approximated by sampling and fitting by successive iterations which refine the model parameters. However, it is observed that in real OSNs, node attachment is also influenced by the "friends of friends" phenomenon, whose consequence is that the probability of attachment is biased to form "triangles".

2.3. Social networks and computers

In this section we look at three aspects of social networks and computers. Firstly we review a brief history of online social network applications. Secondly we review some of the classic and most recent benchmarking datasets used by the SN and OSN analysis community. Thirdly we briefly mention some of the applications and development software used for SN graph analysis.

2.3.1 On-line social network applications

In social network applications, each user is typically defined by a profile, together with a functionality which facilitates searching for and aggregating contacts in a contact list. For each contact to be established, both parties have to mutually accept to create the 'link'. Other functionality is provided such as a ‘chat’, ‘photo albums’ and a ‘wall’ in which the user can publish messages and content which are ‘broadcast’ to the contact list. Online applications, such as games, allow the user to participate, compete and collaborate with other users.

An online social network can be generically understood to be some kind of computer application which facilitates the creation or definition of social relations among people based on acquaintance, general interests, activities, professional interests, family and associative relations, and so on.

Some of the most popular worldwide OSNs are Facebook, Twitter, LinkedIn, Google+ and MySpace, with a number of users ranging from 800 million (in 2011) for Facebook to 61 million for MySpace. Different countries also have their specific applications which are most popular domestically. In the case of China, RenRen (the equivalent of Facebook), has approx. 160 million registered users, of which 31 million are considered active. Weibo (a social microblogging application similar to Twitter) is claimed to have 300 million registered users. Spain has Tuenti, Hi5 is popular in Central and South America, Orkut (India and Brazil), StudiVZ (Germany), and Skyrock (France), among others[11].

Some applications are specific to photo sharing, such as Flickr or Picasa, or videos and music, such as YouTube or Spotify.

A Brief history. Starting in the 1960's, some of the first online networking services, such as Usenet[12], ARPANET and BBS, America Online and CompuServe, already displayed rudimentary OSN features[13].

With the advent of the WWW, in 1994 Geocities[14] became one of the first applications to make use of this new environment to facilitate the interaction between people via chat rooms. Subsequently, in 1997, applications such as SixDegrees[15] incorporated more contemporary OSN functionality to manage "user profiles" and "friend lists". Other notable applications were Friendster in 2002, followed by MySpace and LinkedIn in 2003 [16].

Facebook[17] was launched in 2004, and by 2009 it became the largest social networking site. We could say that the great period of growth in the use of OSN applications manifested itself throughout the years 2005-2010.

2.3.2. The analysis of social network datasets

In this section we distinguish between three types of datasets used by social network analysts: (i) social networks which have been represented in a form which allow them to be analyzed by computer programs; (ii) data logs of computer applications which are not strictly OSN applications, such as email systems, mobile telephone logs, and so on; (iii) data logs of online social networks.

(i) In the first type of dataset, we have, for example, Karate[18] and Dolphins[19], which are small graphs which have been used extensively for benchmarking. Zachary's "Karate club", consists of 34 members of a Karate club, where two key nodes are the club's administrator and the club's instructor. There is a polarization, in more or less equal parts, of the club's members towards these two key members. The Dolphins data set represents a social network of a community of 62 bottlenose dolphins studied by Lusseau et al. in New Zealand [19]. Lusseau et al. compiled the dolphin data from a 7 year field study in which ties between dolphin pairs were established by observation of statistically frequent associations.

In [20] Girvan and Newman conducted empirical tests on different datasets, including "College football", in which the teams are the vertices and the edges are the games between respective teams, a "collaboration network" consisting of 271 vertices which correspond to scientists resident in the Santa Fe Institute over a two year period, and a "food web" consisting of 33 vertices corresponding to an ecosystems principal taxa (species) and in which the edges represent trophic relationships (which taxa eats which).

Many large graphs for benchmarking, some of which have been used extensively in the literature, are available from the SNAP Stanford Large Network Dataset Collection[21]. For example, the dataset cit-HepTh[22], represents relations between citations and cited authors of scientific papers in the high energy physics field, comprising of 27,770 nodes and 352,807 edges; a collaboration network of high energy physics papers, ca-HepTh[23] has 9,877 nodes and 51,971 edges.

(ii) In the second type of graph dataset we have, for example, the Enron dataset[24], made up of a log of emails sent and received between employees of the Enron Corporation during a given time period. This is available online from the SNAP dataset collection, and consists of 36,692 nodes and 367,662 edges. Also, in [25], Seshadri et al. analyze a large dataset of mobile phone calls (one million users and ten million calls). Seshadri examines the distributions of three key factors: phone calls per customer, duration time per customer, and number of distinct calling partners per customer.

(iii) In the third type of graph dataset we have the Epinions ‘Who-trusts-whom’ network[26], consisting of 75,879 nodes and 508,837 edges; The LiveJournal online social network dataset[27] consisting of 4,847,571 nodes and 68,993,773 edges; the wiki-Vote Wikipedia ‘who-votes-on-whom’ network[28] with 7,115 nodes and 103,689 edges; the Flickr images sharing common metadata[29] with 105,938 nodes and 2,316,948 edges; and finally the Twitter dataset[30] made up of 476 million tweets collected between June-Dec 2009, representing 17,069,982 users and 476,553,560 tweets. All of these datasets are available at the SNAP website[21] We also note that some applications such as Twitter[31] and LinkedIn[32] have made APIs (Application Programming Interfaces) available for programmers who wish to perform their own ‘scraping’ of these OSNs, given certain data privacy restrictions.

Finally, as well as the three types of graph data we have mentioned in this section, there is also the synthetic data generated by different models and by different researchers for specific purposes. Through the rest of the survey we will comment examples of synthetic datasets and of course the many other datasets used by researchers in their work.

2.3.3. Applications and software for social network analysis

With respect to software for OSN analysis, on the one hand there are the “off the shelf” applications such as Gephi[33] for visualizing graphs and calculating different graph statistics, including community labelling. Gephi includes as standard the following metrics: node centrality, between-ness, closeness, density, path length, diameter, HITS, modularity, and clustering coefficient. Gephi also has a Java API interface for developers. Another popular application is NetMiner[34] which is a commercial software system with specific modules for Twitter data analysis.

On the other hand there are software development libraries and databases for programmers. ‘Neo4J’ [35] is a graph database software for high performance processing with a Java API for ‘big data’ requirements. The ‘Python NetworkX’ graph library [36] includes generators for classic graphs, random graphs, and synthetic networks, standard graph algorithms, and network structure/analysis measures. JUNG (the Java Universal Network/Graph Framework) [37] is an open source graph modelling and visualization framework written in Java.

Finally, for those who prefer programming in the ‘C’ language, there is the ‘igraph’ library and API [38], and the Stanford Network Analysis Platform (SNAP) [39] is a general purpose, high performance system for analysis and manipulation of large networks, written in C++.

2.4. Graph mining

In this section we cover the chosen key themes of graph mining: ‘classification/topologies’, ‘prediction’, ‘efficiency’, ‘pattern detection’, ‘measurement and metrics’, ‘modelling, evolution and structure’, ‘data processing’, ‘influence and recommendation’ and ‘communities’.

2.4.1. Preamble

Graph Mining can be considered a specialization of Data Mining, the objective of the latter being to process data which is difficult for humans to meaningfully interpret, and identify/extract high value knowledge from the data. For example, a data mining application may analyze a 1 Terabyte database of insurance transactions in order to identify patterns of fraudulent behaviour. The techniques which are used to process the data and extract the knowledge are in general statistical analysis and modelling techniques and/or machine learning methods using artificial intelligence concepts. Thus, we could say that the objective of Graph Mining is similar to that of Data Mining but applied to graphs.

However, graphs have specific properties, especially with respect to the way the data is represented and interrelated, which differentiate it from ‘tabular’ data, and require specialized techniques. In [40], Cook and Holder define graph based data mining as the task of finding novel, useful and understandable graph-theoretic patterns in a graph representation of data.

2.4.2. Prediction/supervised learning

The objective of prediction and supervised learning is to create a data model which is able to learn outcomes from a historical data set, and apply them to a new dataset for which the outcomes are unknown. For example, we can train a model to classify users from three years of historical transactional data, and apply it to classify new users in an online application. This is clearly a wide field, and in this section we will limit the scope to look at two examples of contrasting use. In the first example prediction is used predict link formation between nodes in a social network, and in the second example machine learning is used for identifying common sub-graphs.

Firstly, the prediction of link formation in social networks is considered by Liben-Nowell and Kleinberg in [41]. The authors evaluate different metrics for link-prediction in social networks, that is, how to infer which new links are likely to be created between nodes in the near future. Their approach is based on measures for analysing the "proximity" of nodes in a network, extracted from the network topology alone. A benchmarking is carried out using a battery of different predictor measures for assigning a "score" (Jaccard, SimRank, hitting time, rooted Pagerank, Katz, ...). The test datasets include 'astro-ph', 'cond-mat', 'hep-ph', and 'qr-qc', which represent "co-authorship" networks of authors in different academic disciplines. The relative performance of their method is compared with 'random prediction', 'graph-distance prediction' and 'common-neighbours prediction'. Also the number of common predictions between methods is evaluated. It is found that the best predictor (Katz on the gr-qc dataset) only has a predictive success rate of between 15% and 48%.

Cook and Holder, in [40], discuss the task of identifying frequently occurring sub-graphs in graph transactions, with minimum levels of support. In the context of machine learning, Cook proposes a non-supervised method (hierarchical conceptual clustering). With respect to supervised learning, the problem of choosing adequate examples and counter examples is discussed. Also, "blurred graphs" are proposed as a possible solution for cases which are difficult to categorize, and more generally this could be considered a form of "fuzzy" representation.

2.4.3. Efficiency

As we have already mentioned, computational cost is a key aspect of almost any operation or calculation that we realize on a graph. Many operations, such as calculating the average path length, or looking for isomorphic sub-graphs, are NP-Hard, thus we need to use an efficient data representation and/or heuristic in order to achieve a reasonable execution time and memory usage. Two typical data structures used for representing graph data are adjacency matrices and adjacency lists.

The adjacency matrix is a commonly used solution, in which a graph with N nodes is represented by a N × N matrix, and the ones and zeroes in the cells indicate if two corresponding nodes are connected, or not, respectively. Ramamoorthy's 1966 paper [42] is a key theoretical reference in this field, which considers the analysis of a graph using connectivity considerations. Graphs are considered as representing physical systems represented by weighted graphs, and which have characteristic "generating functions". Some key aspects are defined, such as the identification of "essential" and "inessential" nodes in a graph, testing for "strong connectivity" and sub-graph identification. A matrix representation is used for the graph where a cell with a "1" means that a direct connection exists between the corresponding nodes and a "0" means there is no direct connection. A node *j* in a given sub-graph *T* is considered essential with respect to *T* if it can be *reached* by all other nodes *i* in *T*, and if the terminal node *t* in T can be reached from *j*. A graph is considered *strongly connected* if and only if any node is reachable from any other. The set of all strongly connected sub-graphs is found by constructing the *reachability matrix* for each sub-graph and then checking for non-zero row vectors.

A method which reduces the computational costs of processing an adjacency matrix can be found in [43]. "Colibri" [43] is a method for Low Rank Approximation (LRA) applied to the adjacency matrix of a graph, which the authors claim is as effective as existing methods, such as CUR or CMD, while giving a 100 times reduction in computational cost. LRA is an important tool in graph mining for detecting communities in graphs and also for detecting outliers.

Nair et al. in [44] present a unified approach to learning task-specific bit vector representations for fast nearest neighbour search. This type of search is necessary for applications such as information retrieval and nearest neighbour classification. They propose a learning-to-rank formulation to learn the bit vector representation of the data. The 'LambdaRank' algorithm is used for learning a function that computes a task-specific bit vector from an input data vector. The authors claim that their method outperforms the state-of-the-art nearest neighbour methods they have benchmarked on a number of real world text and image classification and retrieval datasets. The method was shown to be scalable and able to learn a 32-bit representation on 1.46 million training cases in two days.

In Section 2.4.7.1 we will consider the mechanism of 'streams' as an efficient form of data processing within Section 2.4.7 which deals with data processing.

2.4.4. Pattern detection

The detection of different kinds of patterns and structures (specific, similar, outliers, and so on) is a basic task for graph data miners. The earlier work of Gibson, Kleinberg et al. [45] and Kleinberg et al. [46] on detecting 'higher level' patterns such as communities (which we will see later in more detail in Section 2.4.9), contrasts with the 'lower level' patterns such as frequent sub-graphs[47][48] and isomorphisms[49].

A key pattern detection task is that of finding frequent sub-graphs, which also requires an efficient form of navigation through the graph. In [47], Chakrabarti and Faloutsos tested different algorithms and paradigms for finding frequent sub graphs, such as "Apriori-like algorithms", "suboptimal greedy beam search", and "ILP inductive logic programming". The authors propose that navigation in graphs could be done efficiently, for example, by a search guided by the power-law degree distribution. This method would navigate the network via high degree nodes and poll each one to see if it knows the desired information. Alternatively, nodes could be ordered by degree value and a binary search performed for a given degree value. It is stated that some sort of backtracking mechanism is often required in crawling.

Frequent substructure analysis is also considered in [48]. In this paper, Yan and Han define a technique called "gSpan" (graph-based Substructure pattern mining), which discovers frequent substructures without the need for generating candidate solutions. It uses a sparse DFS (Depth First Search) code representation for the graph, which achieves a significant reduction in the computational cost in searching. A lexical value ("x", "y", "z", ...) is assigned to each node and these values are ordered to form a "code", which is used by the DFS process. This is equivalent to the use of a sparse adjacency list representation to store the graphs.

Another work which considers mining frequent substructures from graph data is that of Inokuchi et al. [49], which presents a novel "a priori" based algorithm. The method identifies isomorphisms and is applied to chemical compound data. Basic isomorphism search in a graph has a computational cost which is "NP-Hard", thus the authors propose some improvements in search efficiency by using an "adjacency matrix" to represent the data, and frequent patterns are identified by an extended basket analysis method.

Later, in Section 2.4.7.3 we will consider graph matching in detail, and in Section 2.4.10 we will consider community detection algorithms.

2.4.5. Measurement & metrics

A diversity of metrics exist for measuring, processing and characterizing graphs, the most typical being averaged statistics derived from the degree, clustering coefficient and average path length of the nodes and edges. In [47], Chakrabarti and Faloutsos conduct a review of graph metrics, defining typical graph metrics such as: number of nodes and edges in the graph; degree of each node; average degree for all nodes in graph; *cc*, clustering coefficient for the whole graph; *cc(k)*, clustering coefficient for all nodes of degree *k*; power law exponent; and time/iterations since start of processing (supposing that the graph is being generated/processed by some algorithm). Some key data mining themes considered in [47] are as follows: *(i)* detection of abnormal sub graphs, edges and nodes. In order to do this, a definition has to be made for what are considered 'normal cases'; *(ii)* simulation studies on synthetic graphs generated to be as close as possible to the real equivalent; *(iii)* sampling on large graphs - the smaller graph has to match the patterns of the large graph or it will not be realistic; *(iv)* graph compression - data can be compressed using graph patterns which represent regularities in the data.

The authors comment that typical graph characteristics which occur in naturally occurring graphs are *(a)* power laws, *(b)* small diameters and *(c)* community effects. Power laws can be traditional or can be skewed, for example, as a consequence of the presence of a significant sub-community within the global community. In order to show this, the authors plotted the power law distributions of the 'in-degree' and 'out-degree' for the 'Epinions' and 'click stream' datasets. In the latter case (click stream), the plot showed a skewed effect, given that this data was known a priori to contain a significant sub-community. The authors commented that two of the most common deviations are *exponential cut-offs* and *lognormals*.

As different interpretations of the ‘centrality’ of a node, the following metrics are considered: *(a)* a ‘centrality metric’, in which a high degree for a node implies it is more central; *(b)* degree of indirect neighbours; *(c)* ‘closeness centrality’, defined as the inverse of the average path length of a node to other nodes; *(d)* ‘betweenness centrality’, defined as the number of shortest paths which pass through a node; *(e)* ‘flow centrality’, defined as the number of all paths which pass through a node.

Thus, following the scheme described in [47], *OSN graphs can be characterized by:(i)* power laws (of degree distributions, and other values); *(ii)* small diameters (OSNs ≈ 6); *(iii)* community structure as shown by high clustering coefficients, and other indicators.

As the second reference in this section on metrics, Mislove et al., in [50], define a series of properties for OSNs, such as the power-law distribution, the small-world phenomenon, and the ‘scale-free’ characteristic. Mislove et al. statistically analyse four OSNs (Flickr, YouTube, LiveJournal and Orkut) in terms of these different properties. The authors comment that in OSNs, the 'in degree' of nodes tends to match the 'out degree', that OSN networks contain a densely connected core of high degree nodes, and that this core links small groups of strongly clustered, low degree nodes at the fringes of the network. The ***Power-law*** defines that the probability a node will have degree k is proportional to $k^{-\gamma}$, for large k and $\gamma>1$. ***Scale free networks*** are defined as a class of power-law networks in which the high-degree nodes tend to be connected to other high-degree nodes.

Small world networks are defined as having a small diameter and exhibiting high clustering. Mislove’s analysis focuses on the WCC (Weakly Connected Component), which, it is proposed, is the most "interesting" part of the network. As measures, the correlation of the 'in degree' with the 'out degree' is considered, together with the JDD (joint degree distribution). The JDD indicates the frequency with which nodes of different degrees connect to each other, and is approximated by the degree correlation value. The latter is a mapping between the 'out degree' and the average 'in degree' of all nodes connected to nodes of a given 'out degree'. *Scale free behaviour* is studied, that is, the extent to which the graph has a hub-like core, in which high degree nodes are connected to other high degree nodes.

Mislove also defines the *'Assortativity coefficient'*, as a measure of the likelihood for nodes to connect to other nodes with a similar degree (for example, high with high, medium with medium and low with low). The '*Core*' is defined as follows: (i) a minimal set of nodes which must be necessary for the connectivity of the network; (ii) a set of nodes strongly connected with a relatively small diameter.

It is found, in general (for the four OSNs studied), that the average path length (*apl)* increases sub-logarithmically with the size of the core. In the specific case of Flickr, the overall *apl* was found to be 5.67, of which 3.5 hops involved 10% of the nodes in the core with the highest degrees. This implies that high degree core nodes are within approximately four hops of each other, and the remaining nodes (the majority of the network) are at most a few hops away from the core nodes.

One interesting region of the graph is stated as being a 'tightly clustered fringe', characterized by graph properties of local neighbourhoods outside of the core, and which can be identified by using the clustering coefficient statistic.

Mislove states that the clustering coefficient of social networks is between three and five times larger than their corresponding random graphs, and one order of magnitude greater than random power law graphs. It is proposed that this higher than expected clustering, implying strong local clustering, occurs because links tend to be created as a consequence of mutual introductions between people.

The authors confirm that user groups represent tightly clustered communities of users in the social network. Low degree nodes tend to have low community membership whereas high-degree nodes tend to be members of multiple groups. In general, an OSN is structured into a large number of small, tightly clustered local user communities held together by high degree nodes. Thus, the clustering coefficient is inversely proportional to node degree. It is stated that users in the "core" represent "super nodes" in a two-level hierarchy. In terms of the observed *'temporal invariance'*, although there was a great increase in the size of the Flickr network over time (observed during a five month period in 2007), the overall basic structure stayed similar. Finally, Mislove et al. concluded that OSNs tend to have more symmetrical links and higher levels of local clustering, than networks in general (including the Web).

Given the importance of the ‘closeness centrality’ statistic in identifying key nodes and regions, some work has been done on reducing the computational cost of its calculation. In [51], a fast approximation algorithm for the closeness centrality statistic in weighted graphs was presented. In this paper, Eppstein and Wang present an algorithm designed to process “small world graphs”. For this type of graphs the algorithm estimates the centrality of all vertices with a high probability and with a time cost of *O(m)*, where *m* is the number of edges in the graph.

With respect to degree correlations, in [52], a metric is defined which measures the ‘mean degree of the nearest neighbours of a vertex as a function of the degree k of that vertex’. We recall that vertices of high degree k tend to be connected to others of low degree, and vice versa. Another metric is the variance of the distribution. The high transitivity (clustering), in topological terms means there is a high density of triangles ‘ABC’ in the network. That is, if B has two neighbours A and C, it is likely that A and C are also connected, by virtue of their common relation to B. Another measure is the probability of there being exactly m vertices of degree k in the network, and no vertices of degree greater than k. It is mentioned that the real clustering coefficient of the WWW is 0.11 whereas the expected (modelled) calculation gives a value of 0.048. With reference to community structure in networks, the authors comment that the deviance from expected values can be explained mathematically by a phenomenon coined as ‘bond percolation’, which increases the probability of mutual links. Two probabilities are defined: ‘rm’ is the probability that an individual belongs to ‘m’ groups; and ‘sn’ is the probability that a group contains ‘n’ individuals. The authors develop a generative model for graph networks based on conditional probabilities.

2.4.6 Modelling, evolution and structure

In this section we consider three interrelated aspects: modelling (simulation) of OSN graphs, how they evolve over time, and their structure.

In order to model an OSN graph, we have to understand what are its basic building blocks and characteristics. Robins et al., in [10], define the typical characteristics of a graph as being (i) the distribution of degree frequencies, for which d_k is the number of nodes having degree *k*; (ii) the *q-star*, which is a sub graph of *(q+1)* nodes in which one central node is connected to exactly *q* nodes; (iii) *triangles*; (iv) the *geodesic* between two nodes, defined as the shortest path between them; and (v) the local clustering coefficient of a node. Aspects such as clustering, 'characteristic path length' and 'connectedness' are also mentioned, as well as 'exponential random graph models' and their simulations, which are studied in some detail. A probabilistic formula is given which relates a random graph to an observed graph, in terms of the links defined in the corresponding adjacency matrices. However, in social networks the assumption of independent ties is stated as being generally implausible. As a consequence of this, the authors proposed a "Markov dependencies" model, more specifically a parameterized Markov model in which the parameters are proportional to the frequency of four structures in the graph: single edges, two-stars, three stars and triangles.

Robins et al. [10] conducted a simulation for graph sizes ranging from 30 to 500 nodes. For a 100 node graph, up to 500,000 iterations were necessary to reach a stabilization of the statistical values.

The model statistics used were:

(i) Number of edges
(ii) Number of 2-stars
(iii) Number of 3-stars
(iv) Number of triangles

Aggregate measures (the graph statistics) are then calculated for:

(a) Degree distributions
(b) Geodesic distributions
(c) Clustering coefficient

A difficulty was found in the case of the "degree distributions", given that each sample had its own distribution.

An "energy" value was defined and calculated for the graph at each iteration, the objective being to find the situation in which the energy reached a minimum. A formula was derived for this value, and results were shown in terms of the values for edges, 2 stars, 3 stars, triangles and the clustering coefficient. This was done for two different sampling methods, "Markov random graph sampling" and "Bernoulli sampling".

Robins also studied different types of graph structures, contrasting them with the 'small world' model. For example, in the "long path network” the average path length tends to be much longer and the graph is characterized by “long thin strings of nodes”. A low clustering value was also found for this type of networks.

The authors cite *four key conditions* in order for a small world network to develop:

(i) The individuals seek more than one network partner.

(ii) The costs of maintaining many partners is high, therefore there is a tendency against a multitude of partners. Dunbar’s limit [53] gives a natural cognitive, sociological and anthropological maximum of 150. We also comment this limit and more recent studies in Section 2.4.6.3 of this survey.

(iii) There exists some tendency for network partners to agree about other possible partners, which leads to structural balance and clustering.

(iv) If point **(iii)** is applied in excess this produces cliques with insufficient links between nodes in order to give smaller path lengths. On the other hand, if it is not applied enough there will be insufficient clustering in the network.

Robins et al. also comment another type of graph, the "Caveman graph", which is a sort of "worst case scenario". This graph consists of several fully intra-connected sub graphs in which the sub graphs are not inter-connected. That is, sub graph A is completely disjunctive from sub graph B, also A is disjunctive from C and B is disjunctive from C.

To conclude this first part on models, we consider the R-Mat model defined by Chakrabati et al. in [54]. This model uses a statistical approach and a recursive process. The objective of the research is to model an existing graph of real data, thus deriving its parameterization in terms of given descriptor variables. A typical adjacency matrix of {0,1} values is used to represent the graph (nodes, edges). The authors state that one of the challenges in modelling real graphs, such as social networks, is replicating the power law distributions, skew distributions, and other reported structures, such as the "bow-tie" and the "jellyfish" (in Internet), while maintaining a small diameter for the graph. The computation cost of generating the graph is also an issue. The authors indicate that a model of a social network must also display a "community structure", giving examples such as soccer and automobile enthusiasts, the latter of which can be further subdivided into motorcycle and car enthusiasts. They also consider cross-links between communities which denote persons with diverse interests (e.g. soccer AND automobiles). In order to represent this, a recursive partitioning is carried out, which can be considered as a binomial cascade in two dimensions. The expected number of nodes c_k with out-degree k is given by:

$$c_k = \binom{E}{k} \sum_{i=0}^{n} \binom{n}{i} \left[\rho^{\mathrm{n-i}} (1-\rho)^{\mathrm{i}} \right]^{k} \left[1-\rho^{\mathrm{n-i}} (1-\rho)^{\mathrm{i}} \right]^{E-k} \tag{1}$$

where 2^n is equal to the number of nodes in the R-MAT graph (typically n = $\log_2 N$), ρ is the probability of an edge falling into partition *a* plus the probability of an edge falling into partition *b*, and E is the number of edges in the real graph. The method is tested on two real datasets, "Epinions" and "click stream". Descriptive parameters are used such as degree distributions, number of reachable pairs, number of hops, effective diameter and stress distribution.

2.4.6.1 Evolution: The general consideration of this theme is how the evolution (growth) of social networks can be modelled and empirically measured. A diversity of approaches to the analysis and modelling of evolution in OSN graphs can be found in the literature, ranging from authors who just statistically analyze evolution in real OSN datasets (typically over time), such as Viswanath et al., in [55] and Kossinets and Watts in [56], to authors who study specific aspects and try to ‘model’ the evolution process. With reference to the latter, Tang, et al., in [57] try modelling a ‘multi-mode’ network, that is, one which contains different types of user and actions by those users, whereas Leskovec et al. in [58] define a graph generator called the “forest fire” model which tries to reflect the way link creation propagates through the network. Finally, [59] focuses on the disconnected components of the graph and the incorporation of weights, and the authors propose an improved version of the “forest fire” model.

The theme of community evolution in Dynamic Multi-Mode networks is studied in [57]. The authors propose that an understanding of the structural properties of a network will help in balancing problems and identifying key influential factors. A crucial aspect for modelling evolution is of course the temporal dimension. The authors give a theoretical presentation and pseudo code for modelling a "multi-mode network", that is, a network which has different types of user and actions by those users. The idea is to progressively refine the model using data with ordered time stamps, and weighted attribute values. From this, each user can be assigned to a corresponding community. In order to evaluate the model, different "noise" levels are introduced into synthetic datasets. The model is found to work well for a medium level of noise. In order tune the model's parameters, online clustering and evolutionary multi-mode clustering are applied to the data. Apart from the synthetic data, two real world datasets were tested: *(i)* the Enron email corpus and *(ii)* the DBLP academic publications database. The Enron data is filtered to only include users who send and receive at least 5 emails, which reduces the dataset to 2359 users. Different methods were used to evaluate the results, although it is stated that the true community clusters could not be exactly known *a priori*. One clear observable trend was the evolution over time of each dataset, as shown by each snapshot. One drawback of the method is the requirement of an *a priori* definition of the number of communities and the weights for temporal and interaction information.

The structure and evolution of online social networks can also be analyzed from data logs of applications like Yahoo360 and Flickr. In the study carried out by Kumar, et al. in [60], the authors discovered three regions: *(a)singletons* which do not participate in the network, *(b)* isolated communities which display a dominant star structure, and *(c)* a giant component anchored by a well connected core region. The authors present a simple model which captures these three structural aspects. Their model parameters are *(i)* user type distribution (passive, inviter, linker); *(ii)* preference for giant component over the middle region; *(iii)* edges per time step.

A specific data-log of the Facebook application, corresponding to the New Orleans geographical region, was collected and analyzed by Viswanath, et al., in [55]. Their study focused on the evolution of user interaction, in which it was found that the structural network (links between accepted friends) is not a very true picture of the real friends of an individual, because many of the users (in Facebook) are not very discriminative when they aggregate persons as "friends". Thus, Viswanath proposed that the measure of "activity" will give a much better picture of who communicates with whom, where the intensity of "activity" is proportional to the strength of the relation. However the activity measure used is that of "writes to wall", and many users of Facebook also use others communications channels, such as the chat box, sending an email, and so on. Also the dataset used is skewed with respect to the general Facebook community, because the users were selected by geographical region (New Orleans, USA). However, some useful conclusions and implications can be derived from their study.

An empirical analysis of the evolution of a social network in which the authors collected their own data from a university faculty environment, is described by Kossinets et al. in [56]. The authors constructed their own dataset from emails and other data about students, faculty and staff of a large university. The data covers a one year time period. They use three types of data: *(i)* registry of e-mail interactions - each email message has the timestamp, sender and list of recipients, but not the content; *(ii)* personal attribute information such as status, gender, age, department affiliation, number of years in the community; *(iii)* complete lists of classes attended or taught, for each semester. It was found that the network is influenced by the organizational structure of the environment (the university, in this case), and by the network topology. The authors found that the general network characteristics tend to reach equilibrium and are more or less constant, whereas the individuals are much more volatile. However, we assume that the natural volatility of the student population has skewed the results in this direction. Some key structures looked for were triadic closures, cyclic and focal closures. A multivariate survival analysis was conducted using the following attributes: ‘strong indirect’, ‘classes’, ‘acquaintances’, ‘same age’, and ‘same year’. The effect of gender was studied by comparing pairings of male-male with female-male, and female-female with female-male. It was found that the 'average vertex degree', the 'fractional size of the largest component' and the 'mean shortest path length' all exhibit seasonal changes. On the other hand, the distribution of "tie strength" was found to be stable in the network as a whole over time. Users who were part of bridges also had a tendency to be transient. Although the bridges may act to diffuse information across whole communities, Kossinets et al. found them to be unstable and not permanently represented by particular individuals. It was found that users did not "strategically manipulate" their networks, even though it was technically possible, because there was no motivation. The results are interesting although they have to be considered in the specific context of the study data and environment, and therefore the findings may not necessarily be generalizable to other OSN domains.

A model called "Forest Fire" (with reference to the way link creation propagates), is presented in by Leskovec, et al., in [58]. It has been extensively referenced in the literature, hence we will consider it in some more detail in this Section. In order to define their model, the authors first study four "social network" datasets over time, in order to see how they change with respect to static models. The datasets studied are 'arXiv citation HEP-TH', 'patents citations', 'autonomous systems (internet routers)' and 'affiliation graph (ArXiv)'. The main conclusions are that the graphs tend to get denser over time, and the diameter tends to shrink, this last conclusion going against 'conventional wisdom'. They define a new graph generator, called the "Forest Fire" model, which is defined by the following:

- A densification exponent
- A difficulty constant
- A difficulty function
- The number of nodes and edges at time 't'
- A community branching factor;
- The expected average node out-degree
- The height of the tree
- *H(v, w)* , which is the least common ancestor height of v, w
- The forest fire 'forward burning probability'
- The forest fire 'backward burning probability'
- The ratio of backward and forward burning probability

In terms of structure, the "rich-get-richer" (or preferential attachment) phenomenon is cited as the explanation of the heavy tailed in-degree power-law distribution. Recursive community structures were found for computer networks based on geographical regions. For the patents dataset, the same situation was found in which conceptual groups ("chemistry", "communications", ...) exist. In true OSNs on the other hand, users tend to group together based on "self-similarity". It is noted that in a citation database, a paper only generates outward bound links when it is created. On the other hand, inward bound links will be progressively generated and incremented over time. As a consequence of their observations, the authors require that their model creates a graph with the following characteristics: (i) "rich get richer"; (ii) "copying" which leads to communities; (iii) community guided attachment (densification); (iv) shrinking diameters.

In the basic "Forest Fire" model, two probabilities are used: p, which controls 'forward burning', and r which controls 'backward burning'.

The generative model is as follows. Node v forms out-links to nodes in G_t according to the following process: (i) v first chooses an 'ambassador node' w uniformly at random, and forms a link to w; (ii) a random number x is generated that is binomially distributed with mean $(1\text{-}\ p)^{-1}$. Node v selects x links incident to w, (among links in and out), but selecting in-links with probability r times less than out-links. w_1, w_2, ... w_x are designated as the other ends of the chosen links; (iii) v forms out-links to w_1, w_2, ... w_x and then applies step (ii) recursively. As the process continues, only unvisited nodes are included, thus avoiding cycles. In general, the "burning" of links in the "forest fire" model begins at w, spreads to w_1, w_2, ... w_x, and proceeds recursively until it dies out.

In contrast, a study which focuses on the disconnected components of a graph and the incorporation of weights is that of McGlohon et al. [59]. The following questions are posed: how do the non-giant weakly connected components behave over time? What distributions and patterns are maintained by weighted graphs? Can a generator be produced which models these two behaviours? The following definitions are given: GCC, Giant Connected Component; NLCC, Next-Largest Connected Component; the Diameter as the 90th percentile of the pair wise distance among all reachable pairs of nodes; the weighting scheme of edges can be multi-edge or edge-weights; E(t) is the number of edges over time, N(t) is the number of nodes over time, and W(t) is the total weight of the edges over time.

Ten different datasets are tested, including, arXiv, patents, IMDB (movies), BlogNet, NetTraffic and DBLP. As the result of an empirical analysis over time, the following observations were made: *(i)* real graphs exhibit a "gelling point" at which the diameter spikes and several disconnected components gel into a giant component; *(ii)* after the gelling point, the secondary and tertiary connected components (NLCCs) remain of approximately constant size; *(iii)* a "fortification effect" occurs in which an increase in the number of edges in the E(t) graph gives rise to a total weight W(t) which is super-linear with respect to E(t); *(iv)* the power law distribution is similar at different snapshots over time for the 'in' and 'out' degrees; *(v)* the 'self-similar weight' displays a 'bursty' behaviour over time, with a parametrical fractal dimension; *(vi)* it is possible to calculate an entropy from the 'self-similar weight'. These empirical observations lead the authors to define what they call a "butterfly model", which, it is claimed, is more robust than the "forest fire" model. Its properties are: constant NLCC sizes; densification following the power law; a shrinking diameter (after the gelling point); and power laws for 'in' and 'out' degree distribution. It has parameters, $\mathrm{p}_{\mathrm{host}}$, p_{link} and p_{stepi} , uniformly assigned from [0,1]. In the model, incoming nodes may choose more than one starting point, and a new node k_i has probability p_{host} to select the next starting point *h*. *h* is randomly picked and has probability p_{link} to be linked by k_i, and k_i has probability p_{stepi} to pick one of *h*'s neighbours and continue this process recursively.

In [54], Chakrabarti et al. study the tendency of how OSN networks grow. They observe that new links tend to form on nodes following a power law distribution of their degree (the current number of links they have). The authors comment that the tendency can be defined in sociological terms as "the rich getting richer" (or "cumulative advantage"). The implication is that new nodes will tend to be attracted to form links with existing nodes which have a high degree. They define the in-degree distribution as a power law with exponent given by: $\gamma_{\mathrm{in}} = 1 / (1\text{-}\beta)$. Some mention is given to fractal structures for the geographical distribution of Internet routers.

2.4.6.2 Structure 1: How OSN graphs are structured is related to modelling and to the basic elements we have considered in the previous Section. However, in this Section we will focus on what structure means for the overall graph topology.

The simple definition of graph density is the number of *edges* E divided by the number of *vertices* V. In [61], Randic and deAlba extend this definition by defining what they call 'relative density'. This is defined in terms of the following two metrics: E/E*, where E* is the number of edges in the complete graph having the same number of vertices, and Z/Z* which is defined as the quotient of the number of zeroes Z, divided by the number of ones in the adjacency matrix Z*. A taxonomy of graphs is given with the following categorization: planar/no planar, cyclic, acyclic, transitive, Eulerian, Hamiltonian, bipartite, polyhedral, n-connected, cubic, complete, complete bipartite, isospectral, endspectral, cages, hypercubes, saturated and maximally saturated.

The structure and function of complex networks is considered by Newman in [62]. This study makes an inventory of definitions of metrics and topologies similar to those of [52] and [50], however some distinct definitions are made. For example, a hyperedge is defined as an edge which joins more than two vertices together, and a hypergraph is defined as a graph which contains one or more hyperedges. A bipartite graph is defined as a graph which contains vertices of two distinct types, with edges running only between unlike types. A component is defined as a set of vertices that can be reached from a given node by paths running along edges of the graph. Different kinds of networks are considered: social, informational, technological and biological. Network resilience is considered, and a plot is made of the fraction of vertices removed versus the mean vertex-vertex distance. Community structure is also considered and the dendrogram (hierarchical clustering) is described as a way of identifying communities. The specific domain of epidemiological processes is discussed in the context of the spread of viruses. The SIR and SIS models are mentioned. With reference to network search, Newman proposes using Web keywords or making use of the skewed degree distribution to find results more quickly. ‘Phase transition’ is considered on networks modelled by statistical mechanical models. It was commented that in the limit $n \to \infty$, a model has a finite-temperature transition for all values of the shortcut density $\rho > 0$.

Newman and Park, in [52], consider non-trivial clustering (network transitivity) and positive correlations (assortative mixing) between degrees of adjacent vertices. They comment that social networks are often divided into groups or communities. The "small world" effect is mentioned, together with the skewed degree distributions, and positive degree correlations between adjacent vertices (in most other networks they have negative correlations). They also mention ‘network transitivity’, that is, the propensity for vertex pairs to be connected if they share a mutual neighbour.

One specific graph element of interest is the “bridging node”, which, as the name suggests, acts as a link between different areas of the graph. Hwang, et al., in [63], present and empirically evaluate a metric called "bridging centrality", which the authors propose is highly selective for identifying "bridges" in networks. They identify an impediment of current definitions of bridge metrics which tend to have a broad specificity but a narrow selectivity. Bridges are defined as a sensitive part of a network given that their deletion may produce a major disruption to the entire graph. Hwang also presents a novel graph clustering approach, using the "bridging" points as limits. A "bridge" is defined as a node or an edge which connects "modular" regions in a graph. The clusters are validated by calculating their precision and recall values. They test their methods on the following datasets: synthetic data; two social networks (a "physics collaboration network" and a "school friendship network"); the AT & T Web Network; and a biological network (Yeast Metabolic Network). They state that social networks differ from computer and biological networks in their clustering properties and in that they show positive correlations between degrees of adjacent nodes.

2.4.6.3 Structure 2: One of the characteristics of graph mining which differentiates it from data mining in general is the way in which data records (nodes, edges) are inter-linked. The links (edges) between nodes create a structural dependence and the analysis and search for the structural forms themselves become an objective of graph mining. Online social networks represent a particular type of graphs which have their own peculiarities, such as the “small world phenomena”, and the presence of “cliques”.

An example of the analysis of the topological characteristics of large online social networks is that of Ahn et al. [64]. Ahn et al. define the topological characteristics as a set of different metrics which can be used for measuring/characterizing the graph, but *not* for classifying different sub-graph types. Three different social networks (Cyworld, MySpace and Orkut) are evaluated.

Ahn comments that in social networks, a network of normal friend relations displays different characteristics to a 'clique' based network. Some examples of clique based networks would be: movie actors (similar to a guild type network), scientific collaborators, or members of a dating web. Also, a testimonial (or recommendation) network, such as ‘Cyworld’, is different because it is a closer representation of the real off-line relationships between individuals. Ahn et al. use the typical *metrics* for measuring a graph: degree distribution, clustering coefficient, average path length (also known as degree of separation), and graph diameter. Another metric, the degree correlation (or assortativity) is also defined, which measures the correlation between the degree of a node and the degree of its neighbours. That is, a mapping between a node of degree k and the mean degree of nearest neighbours of those nodes of degree k. If a graph's assortativity is negative this means that hubs (nodes with many links) tend to be connected to non-hubs, and vice versa. It is commented that social networks tend to be assortative whereas other types of networks tend to be disassortative (hubs tend to be connected to other hubs). It is said that this is a unique characteristic of social networks with respect to other types of networks. The results of plotting the power law distribution for the Cyworld network gives a curious result in that two different distributions are found for the users, which the authors propose implies that CyWorld is *two different networks in one*, corresponding to two different types of user: users with testimonials and users without testimonials. Finally, Dunbar's limit [53] is mentioned, which states that on a neurological, sociological and anthropological basis, the maximum theoretical limit of the number of friends of an individual is approximately 150. Recent studies, such as that of Goncalves [65], have validated this limit for OSNs, by analysing Twitter conversation logs. However, the motivation of users for aggregating a greater number (than 150) of 'friends' may due to 'marketing' and 'broadcasting' purposes rather than the intention of one to one interaction.

The "small-world phenomenon" is a characteristic that differentiates OSN graphs from graphs in general. The phenomenon represents the observation that only a small number of connections are necessary to link two nodes in very large OSN graphs. In human relation terms, it means that two people, highly differentiated socio-economically and geographically, are often only a small number of links (on average, six) away from each other in an OSN graph. This phenomenon is studied by Kleinberg in [66], in which an algorithmic perspective is presented in order to analyse and explain why it occurs. Questions are posed such as “Why doesn't this overload/saturate the network?” and “Why the number ‘6’?” With respect to the first question, it is proposed that although a node is potentially just six steps away from any other node, the probability that a given node sends a message/tries to contact a node at distance 6, is very low. With respect to the second question (why the number 6), the author proposes that it has to do with the inverse square law. In order to show this, a formula is derived in terms of powers of two which includes the number six (the maximum number of steps from one node to another), as an upper bound for the inverse-square distribution, thus:

$$[4\log(6n)d(u,v)^2]^{-1} \tag{2}$$

where log is the natural (e) log, *n* is the number of individuals, *u* is a given node, *v* is another (target) node in the graph, and d(*u*, *v*) is the distance between the two.

Finally, in contrast to the particular topology of OSNs, the topology of the WWW has also been studied by different authors, however it is out of the scope of the current survey to enter into details of the structure of the Web. We can briefly state that the topology of the WWW has a distinctive structure, which was first defined by Broder in [67] as looking like a "bowtie", made up of a central “strongly connected” component (SCC), one side of the bow being an ‘IN’ component, the other side the ‘OUT’ component, and with ‘Tendril’ components attached to the IN/OUT components. It is noted that the Web has a very large SCC, and is therefore very resilient to node deletions.

2.4.7. Data processing

In this section we consider three key aspects of processing OSN graphs, which are especially relevant for high volume data: processing the data as a stream, sampling and searching.

2.4.7.1 Streams: One solution to the problem of processing very large graphs is to input the data as a ‘stream’. A *streaming model* is defined as being a data feed in which data is received as a continuous flow, and in which the graph is revealed one edge at a time. One work which considers this approach is that of Feigenbaum et al. in [68], which presents a ‘hybrid’, or semi-streaming model. A semi-streaming model receives the data as a stream but also has available a data space of n × m bits, where n is the number of nodes and m is the number edges, and which acts as a sort of “cache”. The stream may be organized in different ways, for example, if the graph data consists of an adjacency matrix or adjacency list, the edges incident to each vertex can be grouped together. It is clear that specialized algorithms, which act in several passes, are required to calculate metrics such as the shortest path or the network diameter, when the graph data is revealed progressively in a stream.

Another paper which considers how to efficiently process graph data when the data arrives as a data stream is [69]. The authors (3) consider graph streaming in terms of two key parameters: (i) the number *p* of sequential passes over the data, and (ii) the size *s* of the working memory in bits. Another parameter is the ‘per item processing time’, which the authors propose should be kept small. In order to define lower bounds on these values, a trade off is considered between *p* and *s*. A model, called *W-Stream* uses an intermediate temporary stream, which is generated *on the fly*, in order to increase processing efficiency for graph calculations such as "directed shortest path", which can be solved in O((*n* log$^{3/2}$ n) /√*s*) passes, and "undirected graph connectivity", which can be solved in O((*n* log *n*) /*s*) passes. An example is given of a computer with 1GB of available main memory using a trade-off algorithm that runs in *p* = (*n* log *n*)/*s* passes, and which can process a graph with 4 billion vertices and 6 billion edges stored in a 50GB file in less than 16 passes.

The specific problem of estimating the PageRank value of web documents for a stream of graph data is considered in [70] by Das Sarma et al. . The authors state that the overall objective of the streaming model is to use a small amount of memory (preferably sub-linear with respect to the number of nodes *n*) and a smaller number of passes. The specific graph computation considered is the probability distribution after a random walk of length l. The authors state that by applying their algorithm for computing probability distribution on the web-graph, they can estimate the *PageRank p* of any node within a given error margin. The computation cost is O(nM$^{-1/4}$) for space and O(M$^{3/4}$) for passes, in comparison with the standard implementation of the PageRank algorithm which requires *O(n)* space and *O(M)* passes. We observe that an equilibrium has been sought between space and passes. The random walk of length l is modelled as a matrix-vector computation, and the probability distribution is estimated by performing a number (*K*) of random walks. An improvement in the space complexity of the random walk is achieved by calibrating/reinterpreting the accuracy parameter.

2.4.7.2 Sampling: Sampling is another key aspect of processing large graph datasets, when it becomes increasingly difficult to process the graph as a whole due to memory and/or time constraints. Sampling should not be confused with filtering. Filtering eliminates records from the complete dataset according to some criteria, for example, “remove all nodes with degree equal to one”. On the other hand, sampling, tries to maintain the statistical distributions and properties of the original dataset. For example, if 10% of the nodes have degree = 1 in the complete graph, in the sample the same would be true.

Typical sampling techniques which work for tabular data do not tend to give good results for graph data. In this Section, we will see how techniques such as the “Snowball” method are specifically designed for sampling graph data.

Snowball sampling: the original generic algorithm was defined by Goodman[71] and is as follows. The algorithm has two parameters: N, the number of iterations and K, the number of neighbours to be included from each current node at each iteration.

When the algorithms terminates, the whole sample set will consist of the union of the samples S_1 to S_N, which, as defined in the algorithm, are mutually exclusive.

In Fig. 4 we see an example of the application of Snowball sampling, propagating from three initial seed nodes (square shaped nodes labelled with S). The number of nodes to be sampled at each iteration (K) is set to three and the number of iterations (N) is set to two. The nodes chosen during iteration 1 are labelled I_1 (circular shaped) and the nodes chosen during iteration 2 are labelled I_2 (triangular shaped). We observe that some of the neighbours of two of the seed nodes are unchosen, given that only up to K neighbouring nodes can be chosen for any given sample node. Also, we see that a region of the graph was not sampled, because there was no seed node in that region.

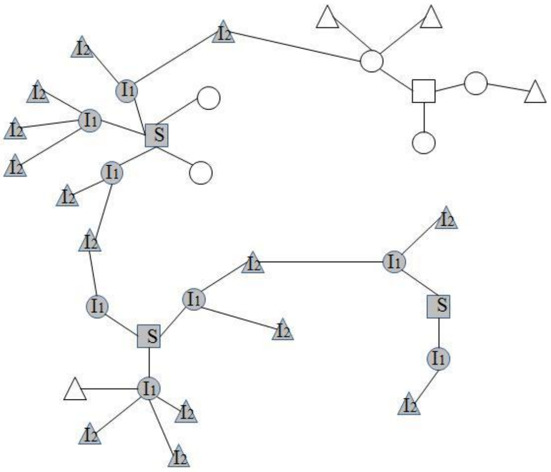


Fig. 4. Snowball sampling: example propagation from three initial seed nodes.

```
-------------------------------------------------------------
                 Snowball Sampler Algorithm
-------------------------------------------------------------
Initialization: a random sample $S_R$ of individuals is drawn from a given finite population.

For each iteration $I=1$ to $N$:
Do
    Each individual $i$ in the current sample $S_I$ is asked to choose $k$ different individuals in the population. The $k$ individuals chosen would typically be $i$'s "k best friends," or the "k individuals with whom $i$ most frequently associates," or the "k individuals whose opinions Si most frequently seeks," and so on.
    It is assumed that $i$ cannot include him/herself in the list of $k$ individuals chosen by $i$.
    If this is the first iteration:
        The individuals who were not in the random sample $S_R$ but were named by individuals in it will form the next sample $S_{I+1}$.
    Else:
        The individuals who were not in the random sample $S_R$ or any other previous sample $S_1 \ldots S_{I-1}$, but were named by individuals in it will form the next sample $S_{I+1}$.
Efor
-------------------------------------------------------------
```

The simplest interpretation of this algorithm in the context of an OSN graph is that a given node will choose its immediate neighbours as the individuals to be included in the next sample. Likewise, those neighbours will choose their respective immediate neighbours, and so on. The result will be a “rippling out” from the original seed nodes, one hop at a time.

A crucial aspect of the Snowball algorithm is the way of choosing the initial seed nodes. In the original version, the choice is random, which would be statistically correct. However, by inspection this may favour the inclusion of a disproportionate number of nodes which are neighbours of high degree nodes. That is, nodes with higher degrees will consume all of their quota (K neighbours) whereas nodes whose degree is less than K will consume less than their potential quota. This may result in a skew of the final distribution of the degree of the nodes in the complete sample. In the following we describe the work of some authors who have considered these problems and their proposed solutions.

Two sampling methods are compared in [54]: *(i)* a full graph data collection and *(ii)* the Snowball method. The latter is implemented by taking well connected seed nodes and growing a graph around them. However the authors confirm the general consensus in the literature that although ‘snowballing’ is an adequate technique for graph sampling, it tends to miss out isolated individuals and include a disproportionate number of neighbours of high degree nodes. In order to solve these problems, the authors propose a random or probabilistically weighted selection of seeds.

BFS and DFS are two common algorithms for crawling graphs, however, it is often not computationally feasible to use them to crawl a complete graph, and therefore sampling methods must be used. In [50], Mislove et al. conduct empirical sampling and crawling tests on datasets derived from Flickr, LiveJournal and YouTube. The Snowball method is defined as being a technique for crawling a graph by the 'early termination of a BFS'. However, the authors confirm that the method tends to overestimate the node degree and underestimate the level of symmetry. Mislove et al. also found that the Snowball method often missed out isolated nodes, the majority of which were characterized as having a low degree and being members of small, isolated clusters.

One possible way to improve the Snowball sampling method is by assigning weights which influence the selection of vertices. This approach is presented by Snijders in [72], which states that without weights, the Snowball method tends to bias the sample giving preference to vertices with a high number of connections. In order to smooth this, a general solution is adopted which weights the selection of vertices. More specifically, one approach is to include only symmetric relations (although for social networks this could be a problem), and another approach is to estimate the frequency of characteristics such as "transitivity", which is defined as the number of chains of length two, divided by the number of triangles. Another consideration is the randomness of the initial sample, which, in order to be evaluated, ideally requires auxiliary external information about the network.

In a more recent evaluation, Shafie [73] considers design estimators for Snowball sampling. Again it is stated that assigning equal weighting causes a heavy bias on high degree vectors, therefore Shafie recommends that sampled elements be weighted by the reciprocal of their selection properties (that is, their degree value). Four weighting schemes are tested, and the one found to be the optimum defines the initial vertex with an inclusion probability of $\frac{1}{N}$, where N is the number of vertices in the graph. Successive vertices will then have an inclusion probability of $d_i \,/\, \sum_{i}^{N} d_i$, where d_i is the degree of vertex i. Hence, the sample selection probability for each possible sample is approximately inversely proportional to:

$$\left[\frac{1}{N} + (n-1) \frac{d_i}{\sum_{i}^{N} d_i} \right] \tag{3}$$

The Snowball sampling is carried out in 'waves', in which the sampling size of wave w_j depends on the degrees of the vertices selected in the previous wave w_{j-1} . Shafie executed a total of five 'waves'. It was concluded that the weightings made a significant improvement, but only for the initial waves, which is when the selection bias has most impact. The best weighting scheme used the observed mean degrees of the samples obtained to estimate the inclusion probabilities. However, in the empirical evaluation, only two synthetic datasets were tested, the first with two equally sized population groups and the second with two unequally sized population groups. It was suggested that the MSE (mean square error) and variance should be used for larger simulations.

A key aspect in sampling is the choice of the initial starting nodes (or ‘seeds’) for extracting the sample. Another aspect is how to measure the ‘quality’ of the derived sample. Both these aspects are considered by Bartz, et al. in [74]. In the first part of the paper Bartz conducts an empirical study of the geometry of three real OSN networks, in which the nodes represent 'prisoners', 'assembly line workers' and 'karate club members', respectively. Structures such as 'terminal nodes', 'two-stars' and 'triangles', are especially studied. Two different graph generator models are defined in terms of the frequency of these structures: **(i)** an exponential random graph model, and **(ii)** a maximum likelihood estimation (MLE) model. The latter part of the paper then deals with sampling. The quality of the model was plotted with the "triangle parameter" on the *x-axis* and the "two-star parameter" on the *y-axis*. Two sampling methods were tried: "multiple bridge sampling" and "Snowball sampling". In the case of the first method, convergence problems were found, MLE being used as the accuracy measure. A critical aspect was found to be the selection of S starting points, which are chosen as one MPLE (Maximum Pseudo-Likelihood Estimator, calculated with multiple bridge sampling using the "two-star", "triangle" and "edge" values as coefficients) and S-1 SMLE's (Snowball method which uses MLE, fitted to the subsample of n/2 nodes). It was found that the estimators for the initial vertices helped convergence when placed in the initial mix, ensuring a more accurate first step. However, the convergence was still considered as sub-optimal, given that between 2% and 4% of the tests did not converge.

The solution to obtaining a good sample from a graph resides in achieving that its distribution fits, or is representative of, the whole graph. One way of doing this is by obtaining *a priori* knowledge about a graph’s structure which can be used by the sampling process. An approach which tries this is found in [75]. In this paper Snijders considers synthetic graph generators (specifically, exponential random graph generators) which induce sub graphs from a complete graph with given statistical distributions. The idea is to elicit knowledge about the way these graphs are generated and apply it to sampling, specifically for the case of the Snowball technique. A parameterized formula was given for the distribution of an induced sub graph (which represents a "weak condition"), and for which a proof was given by mathematical induction. The authors stated that one key problem in Snowball sampling is the leaving out of elements from the sample which are "loosely connected", and which are outside the giant component. Examples of such elements are isolated nodes, isolated dyads (edges), isolated two-stars and isolated triangles. It was stated that the total number of these small structures is dependent on the parameters that determine how larger structures are formed and the connectivity between the smaller and larger structures. It was assumed that mutually disconnected parts of the graph are independent. The analogy was made between trying to obtain a correct distribution from different random samples, and how arbitrary node deletion would affect the distribution of the whole graph. It was said that in the case of unconnected sub graphs, the Snowball sample would be a truer representation of those sub graphs. However, it was also observed that the Snowball method tends to influence, or place certain conditions on, the results.

2.4.7.3 Search and matching: Once we have decided on how to represent and obtain our data, search is the next activity we will probably do. The obvious problem in large graphs is the computational cost. Search can be performed for different motives, for example, (i) in order to calculate some graph metric such as shortest path or clustering coefficient, or (ii) to find a given pattern or substructure.

Dijkstra's classic paper of graph theory literature [76], presents an efficient algorithm for finding the shortest path length between *n* nodes. From the shortest path length between all node pairs, the average shortest path length for the whole graph can be calculated. More recently, in [77] a shortest path searching method was presented with a heuristic for limiting the search area. The objective of the method was to improve on the performance of Dijkstra's basic algorithm while maintaining accuracy, by exploiting the spatial characteristics of networks. The test data consisted of a road network graph, and the empirical test made a correlation between the shortest path and the Euclidean distance.

A favourite computational cost benchmark is that of finding the number of triangles in a graph, where a triangle is defined by three nodes in which each node is connected to the other two. In [78], the problem of counting triangles in graphs is considered, and two space bound algorithms are presented which process an undirected graph defined as a stream of edges. The first algorithm assumes an unordered stream and the second assumes that the stream is ordered such that all edges incident to the same vertex appear consecutively. The algorithms calculate an approximation of 1+ ε with a probability of 1 - δ, where ε and δ are application specific. The authors claim that their method offers a significant improvement in space usage. In another study of computational cost reduction, [79] considers the efficient identification of 'centre-piece sub graphs'. That is, given Q nodes in an OSN, how can we find the node that is the 'centrepiece', that is, a node having direct or indirect connections to all or most other Q – 1 nodes in the graph? The authors apply their algorithm to an authorship network and a DBLP dataset.

Isomorphic matching of graphs: Recently, new efficient matching algorithms for isomorphic matching have become available, such as VF[80]. We recall from the definition of Section 2.1 that two graphs which are isomorphic not only have the same topological structure but also have a one to one mapping of all their corresponding labelled nodes.

VF[80] is based on a depth-first search strategy and uses a set of rules to efficiently prune the search tree. An improved version of VF, called VF2[81], uses more effective data structures in order to further reduce the computational cost of matching. VF2 has subsequently become widely used in the graph mining community. We show the pseudo-code of the matcher is as follows:

```
----------------------------------------------------------------
VF2 Isomorphism Matcher Algorithm [81]
----------------------------------------------------------------
PROCEDURE Match(s)
INPUT: an intermediate state s;
          the initial state $s_0$ has $M(s_0)=\varnothing$
OUTPUT: the mappings between the two graphs

IF M(s) covers all the nodes of $G_2$ THEN
   OUTPUT M(s)
ELSE
   Compute the set P(s) of the pairs candidate for
   inclusion in M(s)
   FOREACH (n, m)$\in$(s)
      IF F(s, n, m) THEN
         Compute the state s' obtained by adding
         (n, m) to M(s)
         CALL Match(s')
      END IF
   END FOREACH
Restore data structures
END IF
END PROCEDURE
----------------------------------------------------------------
```

Washio and Motoda, in [82], consider some theoretical definitions (sub-graph isomorphism, graph invariants, mining measures), and approaches of graph mining which derive from the AI and Data Mining fields. The graph mining approaches are: greedy search, inductive database approach, mathematical graph theory based approach, and the kernel function based approach. They comment that two frequent application domains are the analysis of chemical compounds for medical problems (such as carcinogenesis prediction of chemical compounds), and the pharmaceutical industry's focus on drug discovery. Although these are not OSN graphs, the algorithmic work done in these fields has a direct impact on the data mining of OSN graphs. One specific activity on these application domains is the search for similar patterns/sub-graphs, using nodes which are labelled. A feature space is constructed as a label sequence, and a count is maintained of the number of times a particular vertex appears in a sub-graph. However, for simply anonymized social network graphs, we would have to consider which aspects of these approaches are useful and/or applicable. With reference to **sub-graph isomorphisms**, a graph G is represented as G(V, E, f) where V is a set of vertices, E a set of edges connecting some vertex pairs in V, and f a mapping f : $E \rightarrow V \times V$ between edges and vertices, for example, f(e_1) = (v_1, v_2). Thus, two sub-graphs will be isomorphic if their respective graphs, represented by G(V, E, f), are identical. Graph invariants are defined as quantities which characterize the topological structure of a graph. An important observation is that if two graphs are isomorphic then their invariants will be identical, but the reverse property does not hold. Therefore, a search based on graph invariants would find similar graphs, but not necessarily isomorphic ones. Graph invariants can be the number of vertices, the degree of each vertex, and the number of cyclic loops. An initial search based on graph invariants can be used to reduce the search space for a posterior search for isomorphisms. The following mining measure is defined as "support". Given a graph data set D, the support of the sub-graph Gs, sup(Gs) is:

$$\sup(Gs) = \frac{N^\circ\_of\_graphs\_including\_Gs\_in\_D}{Total\_N^\circ\_of\_graphs\_in\_D} \quad (4)$$

Different search methods are commented, such as heuristic, complete, direct and indirect matching. *Direct matching* finds isomorphisms while *indirect matching* uses some similarity measure to find similar sub-graphs.

Other similarity measures for graph matching: an alternative approach to the topological mapping of all nodes and edges is to use some sort of distance measure based on graph statistics such as degree, clustering coefficient and number of edges. The latter approach, although an approximation of the former, has the advantage of a potentially significantly lower computational cost. For example, consider the following metric which calculates the distance between the sub-graph formed by the immediate neighbours of a given reference node N_1, with the sub-graph formed by the immediate neighbours of another reference node N_2.

$$\begin{aligned} &\text{distance}(A, B) = \\ &\Delta\,(D_R(A), D_R(B)) \times \alpha + \\ &\Delta\,(N_E(A), N_E(B)) \times \beta\ + \\ &\Delta\,(CC(A), CC(B)) \times \chi\ + \\ &\Delta\,(AD_{AN}(A), AD_{AN}(B)) \times \delta\ \square \\ &\Delta\,(SD_{AN}(A),\ SD_{AN}(B)) \times \varepsilon \end{aligned} \quad (5)$$

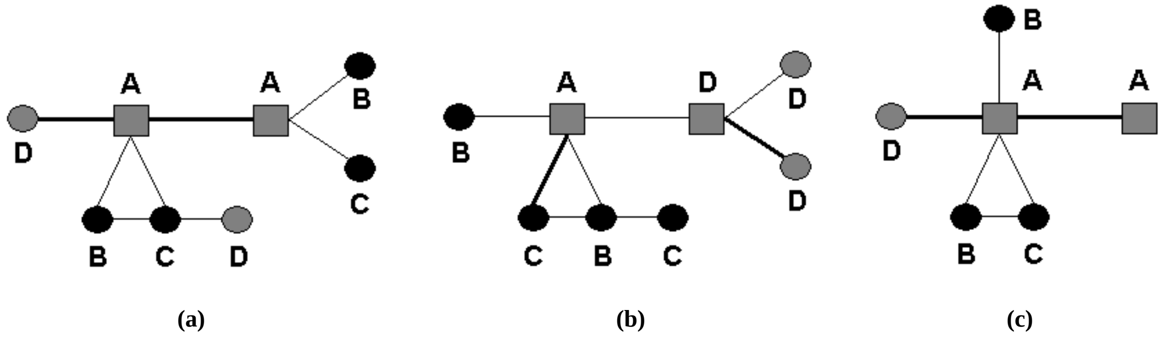


Fig. 5. Three similar graphs, with additional information indicated by labels, shapes and edge activity (shown by thickness of line)

The weight vector {α, β, χ, δ, ε} can be calibrated using a suitable supervised learning technique and target values.

This similarity metric calculates a distance based on sub-graph characteristics which can be pre-calculated. The sub-graph characteristics are: degree of the reference node D_R; number of edges in the immediate neighbourhood sub-graph N_E, clustering coefficient CC, normalized average degree of adjacent nodes AD_{AN}, normalized standard deviation of degree of adjacent nodes SD_{AN}. The first three characteristics are designed to reflect the internal structure of the sub-graph, whereas the last two characteristics reflect a key aspect of the neighbours (their degree), which effectively considers the neighbourhood one hop further out (with the reference node as 'ground zero'). We observe that in order to perform the calculation, all values are normalized against the maximum and minimum corresponding values in the complete graph.

Finally, we can consider additional information when matching a graph, other than its structure. One type of additional information could be the activity registered along the edges, such as the number of Tweets, emails, call durations, etcetera. A second type could be external information about each node/user, such as age, gender, likes/dislikes, geo-location, and so on.

In Figs. 5 we see a depiction of three sub-graphs. At first glance it would appear that in purely topological terms, graphs 5a and 5b are the most similar. However, if we also take into consideration additional information, such as the node labels (A, B, C, …), node shapes and edge activity (indicated by thickness of line) we observe that it is graph 5c which corresponds the best with graph 5a in these terms.

If we use Formula 5, with weights set to 1 to calculate the respective distances between the three graphs of Fig. 5, the results are: distance(a, b) = 0, distance(a, c) = 5, distance(b, c) = 5. The five descriptive factor vectors are {4, 5, 0.33rec, 2.25, 0.957} for graph 5a, {4, 5, 0.33rec, 2.25, 0.957} for graph 5b and {5, 6, 0.2, 1.4, 0.547} for graph 5c. We note that in order to calculate the distance, the values are normalized against the maximum and minimum values for all graphs. For example, for the reference node, the min degree value is 4 (graphs 5a and 5b) and the maximum degree value is 5 (graph 5c). Hence the normalized degree values of the reference node would be 0, 0 and 1 for graphs 5a, 5b and 5c, respectively.

2.4.8 Influence and recommendation

Which persons are the most influential in an OSN? Which Internet Web pages are most influential in a given topic? These are questions which can clearly be of commercial interest as well as being key aspects which help us understand interactions and information flow in an OSN graph.

In this section we refer to three key papers which consider, respectively, authoritative sources in a Hyperlinked web environment [83], finding 'influential' individuals [84][24] and how to maximize the spread of influence through a social network[84].

In the context of Web pages, [83] considers the theme of authoritative sources in a Hyperlinked environment. This paper is related to Kleinberg's work with co-author Gibson on the HITS algorithm [45], which defined and considered Hubs and Authorities. An iterative algorithm for computing hubs and authorities was defined. Also, in the social network context, the concepts of "standing", "impact" and "influence" were defined theoretically. In the scientific citations domain, the "impact factor" was considered, and an improvement was proposed, called the "influence weight". The "influence weight" considers that a journal is "influential" if it is heavily cited by other influential journals which are defined in the same manner, and so on, recursively. The authors stated that overall, their algorithm finds the most densely linked collection of hubs and authorities in a sub-graph G_σ defined by a query string σ. Other concepts were considered, such as "diffusion" and "generalization".

In the context of finding ‘influential’ individuals (that is, individuals who are able to influence other individuals in a social network), Kempe et al. [84] consider how to maximize the spread of influence through a social network. The initial problem is defined as being how to choose a subset of N individuals from the whole graph, who if they were to make a decision, for example, if to buy a given product/service, would influence the maximum number of other individuals in the whole graph to do the same thing. Finding these individuals stated as being NP-hard, and in order to reduce the computation search cost, the "degree centrality" and "distance centrality" metrics were proposed as search heuristics. Two diffusion models, the ‘Linear Threshold’ and ‘Independent Cascade’ are considered. These models represent two different approaches to solve the influence maximization problem. All empirical tests are carried out using the ArXiv high energy physics citation dataset, which once filtered, gave 10748 nodes and 26500 edges. The high degree heuristic (based on node centrality) chooses nodes *v* in order of decreasing degree size *d*. The centrality measure assumes that a node with short paths to other nodes in a network will have a higher chance of influencing them. Four variants were tried for finding the most influential nodes: 'greedy search', 'high degree', 'degree centrality' and 'random'. 'Greedy' combined with 'degree centrality' was found to give the best results, whereas 'degree centrality' on its own worked well to choose the first node but after that showed little or no improvement. This is because the first node tends to be connected to other candidate nodes. However, combining 'degree centrality' with a 'greedy search', in which the already chosen nodes were excluded as candidates, gave significantly better results. All the search methods were applied using a 'weighted cascade model'.

Another study which focuses on finding influential individuals in an OSN is that of Shetty and Adibi [24]. Their approach, in which they analyzed the Enron email database, proposes a "graph entropy" method for discovering important nodes. We recall that the Enron email database contains emails from, among others, top executives of the Enron Corporation. Three types of users are defined: *leaders*, *middlemen* and *followers*. For "graph entropy", they have used Körner's definition [85], which is as follows: a function *H(G, P)* is associated with a graph *G* and an arbitrary probability distribution *P* on its vertex set:

$$H(G,P) = \min_{X \in Y \in S(G), I'x = I'} I(X \wedge Y) \tag{6}$$

where *S(G)* represents the family of stable sets of vertices in G , and a subset of the vertex set is stable if it does not contain any edge. If the mutual information I(X ∧ Y) measures the degree of independence between the random variables X and Y, then the graph entropy will measure the level of independence of the stable sets of G with respect to the vertices.

Shetty and Adibi examined each email to see if it was similar to other emails received by a given individual, and if one transaction was recent with respect to another for a given time window. The basic idea was to measure the change in entropy of the whole graph when a given node was deleted from it. The authors applied their method to the Enron data and it successfully identified the key nodes (employees), benchmarking against the "betweenness centrality" metric. The latter method found nodes in the centre of the graph but not those with the highest authorities, whereas the entropy method did. For example, the key user/node "Louise Kitchen" was only ranked fifth by the "betweenness centrality" measure but second by the entropy model. One clear difficulty of generalizing the entropy model method as defined by Shetty and Adibi to detect important nodes in graphs, is the computational cost. Each node in the network has to be tested individually, that is, dropped from the network, and the entropy recalculated.

The distributions of the number of emails per user and the number of emails sent over time were plotted. The emails were also compared for similarity using Jaccard's Algorithm [86] (also known as Jaccard's Index or Similarity) on the text bodies.

2.4.9 *Community identification: early work and communities in the Web*

In this Section we take general look at initial approaches of defining and analysing community structure, and some specific applications for the Web.

One key structure in an OSN, and indeed in any social network, is the community. Humans are social creatures and tend to aggregate into subgroups based on similar characteristics, shared interests, geographical proximity, and so on. However, given a seemingly chaotic OSN represented as a graph, how do we go about identifying the community structure?

The study of community structure in social networks has been of interest for many years, by multidisciplinary researchers such as Zachary [18], and Freeman [87]. More recently, with the advent of online social networks (Facebook, Twitter, etc.) research in this area has been given a great impulse due to the availability of (some) of this online data for analysis [23][24][47][54][55][88][89][90]. Authors such as Kleinberg[91] and Kumar et al. [60], among others, 'set the stage' for research in this area.

Chakrabarti and Faloutsos, in [54], comment that the clique is the classic community structure, for which there are several "relaxed" definitions, including accessibility criteria such as "at least K nodes" or "at most N hops". Several variants of cliques are proposed, such as "clan", "hops", "core" and "plex". The equivalence of nodes is expressed as "structural", "automorphic" or "regular". The concept of *'social capital'* is proposed in order to define what is to be considered a "well connected node" in an OSN. This may be a node which occupies a special place in the graph, for example, which allows it to broker information or facilitate the work of others. A special node may exploit "structural holes" in the graph or be connected into a dense sub graph. A *'key player'* node can be identified as a node whose removal maximizes the disruption to the network, or a node which is maximally connected to the rest of the network. Some interesting OSN datasets that [54] considers in this context are: "who-trusts-whom" (Epinions), "who-reads-whose-weblog" (Blogspace), and "who-knows-whom" (Friendster).

Newman, in [62], defines and tests three different algorithms derived from the "*shortest path betweenness*" metric. The idea of "*shortest path betweenness*" is simply a counter for each edge E which registers how many shortest paths (to/from nodes in the graph), pass along E. Thus, it is a sort of measure of importance of the edge in the network: if we remove edge E what impact will that have on the nodes in the graph. One difficulty, as always in graph processing, is the computational cost. Complete graph partitioning is NP-complete and runs in time O(n^3) for sparse graphs. *Social networks are considered as having a "community structure"* made up of a number of different communities with dense internal links, which are connected by lower density "inter-community" links. Different analysis methods are considered, such as *hierarchical trees (dendrograms)* and *agglomerative clustering*. It is said that the dendrograms are more successful in representing intra and inter community links, whereas agglomerative clustering tends to lose the inter type links. Three algorithms are tested: (i) 'shortest path betweenness'; (ii) 'random walk betweenness'; and (iii) 'resistor networks'. The random walk is similar to the shortest path measure except that instead of counting the shortest paths that traverse an edge, it counts the number of random walks that traverse the edge. The resistor network is based on electrical theory and applies a "resistance" value to each edge, along which the "current" flows according to Kirchhoff's laws. Then the 'random walk betweenness' algorithm is applied to each edge. A metric is defined for measuring the quality of the solution found, which is based on assortative mixing, and derived in terms of the fraction of edges in the network that connect vertices in the same communities. If this value is high, it implies a good division of communities. The following test datasets are used: (i) an artificially generated network consisting of 128 nodes which contain four communities of 32 nodes each; (ii) Zachary's karate club [18], which contains 34 nodes, two of which are hub nodes; (iii) a collaboration network of scientists (Physics E-print Archive at arxiv.org), consisting of 145 scientists in the largest component of the network, and the remaining 90 scientists belong to smaller components; (iv) a social network of a community of 62 bottlenose dolphins studied by Lusseau et al. in New Zealand [19].

Lusseau et al. compiled the dolphin data from a 7 year field study in which ties between dolphin pairs were established by observation of statistically frequent associations. The network splits into two main groups, and the larger of these groups subdivides into four smaller subgroups. It was found that the dolphin community behaves in a similar manner to a human community, fragmenting when certain key individuals are lost. With respect to the algorithms tested, the 'random walk' and 'shortest path' gave similar results, however the computational cost of both is high and in the case of 'random sampling', the 'betweenness' characteristic is often lost.

One specific area of interest is the study of hyperlink topologies in Internet in order to infer “communities”. Gibson et al., in [45], used the HITS algorithm in order to do this, HITS being one of the most well known algorithms for Internet ranking. It considers two types of important pages: “hubs” and “authorities”. A hub is a page which points to many “authoritative” pages, whereas an authoritative page is one which is pointed to by many hubs. A “good” hub is one which points to many “good” authoritative pages, and a “good” authoritative page is one which is pointed to by many “good” hubs. Thus, a community should contain each of these special page types.

How communities emerge in the Web is also evaluated by Gibson et al. in [45]. Questions are posed such as: ‘how does this evolution depend on the “root set” of sources?’ and ‘how long does it take in temporal terms for a community to form?’ Six topics for generating communities are evaluated: “Harvard”, “cryptography”, “English literature”, “skiing”, “optimization” and “operations research”. The HITS algorithm is applied to form communities based on these themes. The resulting community structures are evaluated in terms of (i) robustness; (ii) topic generalization; (iii) topic hierarchy (tree); and (iv) temporal aspects. A community is considered 'robust' if it doesn't change even though the initial root set is altered. Temporal aspects are considered, such as the definition of a time period over which the community changes significantly. The “core” of the community is defined as the part which does not change over a given time period, whereas other parts of the community are transient. In order to analyze the factors which influence a topic, the author proposes that the HITS algorithm, with its use of eigenvectors, can be used to discover multiple communities associated with a given topic. For example, a single principal and an arbitrary number of sparser non-principal communities could be identified. Empirical tests: for each topic, tests were made for root sizes of 25, 50, 100 and 200. It was found that between 3 and 50 iterations (of HITS) gave similar results, in terms of overlap of the topic community with respect to the full community. A fairly lineal increase (of overlap) was shown from 0 to 10% for an initial root set size of 25, up to a maximum of 20% for a root set size of 200.

Following on from [45], in [46] Kleinberg et al. defined a new algorithm to "trawl" the web for cyber-communities. The “trawling algorithm” differentiates from HITS in that the latter searches for high-quality pages about a specific topic, whereas the “trawling algorithm” is designed to search for N defined topics. It does this by identifying "bipartite cliques" and "bipartite cores", where a "bipartite clique" $K_{i,j}$ is defined as a graph in which every one of i nodes has an edge directed to each of j nodes, and a "bipartite core" C_{ij} is defined as a graph on $i + j$ nodes that contains at least one K_{ij} as a sub graph. However, it is stated that the computational cost of a simple algorithm to search for these structures is not practicable. Thus Kleinberg proposes a two step algorithm: *(step i)* elimination and *(step ii)* generation, consisting of sub steps of successive passes and sorts on the data. The graph is stored as a set of binary relations. To gain speed, two aspects are exploited: *(i)* the fact that the in/out degree of each node drops monotonically after each step; and *(ii)*, nodes are eliminated or included as they are respectively excluded or included as belonging to a core. Other sub graphs stated as being of interest are ‘bidirectional stars’, cliques and directed trees.

Community structure: a community is stated as being a set of nodes where each node is closer to the other nodes in the community than to nodes outside it, a typical measure being the clustering coefficient. Transitivity occurs *iff* triangles exist in the graph, that is, connected triples. The authors propose that communities can be extracted from graphs by two main methods; *(a)* bottom up construction starting with individual nodes and constructing successive hierarchies, using a method such as Dendrograms; *(b)* top down analysis by identifying edges with a high "betweenness" and deleting them. One difficulty is defined as how to choose the initial seed nodes. As a possible solution, the authors propose the HITS algorithm, which uses hub and authority nodes to bootstrap the algorithm. Another consideration is the resilience of a graph, that is, the effect on the graph of removing nodes from it. Other metrics commented are the 'in/out degree correlation', the 'average neighbour degree', and the 'neighbour degree correlation'.

2.*4.10 Communities in OSNs: identification and extraction*

In this section we will consider two specific algorithms for automatic community extraction from a complete graph: that of Newman and Girvan[88], and that of Blondel et al. [89]. Newman and Girvan’s algorithm was effective but slow, whereas that of Blondel et al., designated as the Louvain Method[89], was developed four years later and is much more efficient in computational terms, having now become an ‘industry standard’. In this Section we will briefly describe both algorithms, and discuss some of the results of community extraction for two benchmark datasets.

2.4.10.1 Newman & Girvan’s community search algorithm – implementation details: In this section we describe the high level algorithm for the search and evaluation of communities on complex networks, based on the method detailed in [88].

Newman and Girvan’s algorithm [88] focuses on how to extract a community structure from social network graph data. Two main approaches are defined: (i) the identification of groups around a prototypic nucleus defined in terms of the 'most central' edges, an adjacency matrix being used as the basis to calculate the weights; and (ii) the identification of groups by their boundaries, using the least central edges (frontiers). This second metric is also referred to as "edge betweenness", and is based on Freeman's "betweenness centrality measure" [87]. A summary of the algorithm is as follows: (a) calculate the betweenness for all edges in the graph; (b) remove the edge with the highest betweenness; (c) recalculate betweennesses for all edges affected by the removal; (d) repeat from step (b) until no edges remain.

In order to understand the community extraction algorithms, we first explain what ‘modularity’ is, a measure used during community extraction to measure the ‘goodness’ of the current partitioning.

Modularity[88]: During the elicitation process, the graph is successively divided in components, and the correctness of the community partitions is measured. The quality metric used for a given community is called the modularity. For a graph divided into *k* communities, a symmetrical matrix *e* is defined of order *k2* whose elements $\boldsymbol{e_{ij}}$ are the subset of edges from the total graph which connect the nodes of communities i and j.

The trace of matrix e, denoted as $Tr\ e= \sum_i e_{ii}$ gives the fraction of edges in the graph which connect nodes of the same community. Hence, a good division in communities should obtain a high value for the trace of matrix e. However, this value alone is not sufficient as a good quality indicator, given that if all the edges are placed in the same community this would give the maximum value, $Tr\ e = 1$, but without having created any useful structure. Thus it is necessary to define the sum of the rows $a_i=\sum_j e_{ij}$, which represents the fraction of edges which connect nodes of community i. Following on from this, the modularity metric was defined as:

$$Q=\sum_i \left(e_{ii}-a_i^2\right)=\mathrm{Tr}\ e-\|e^2\| \quad (7)$$

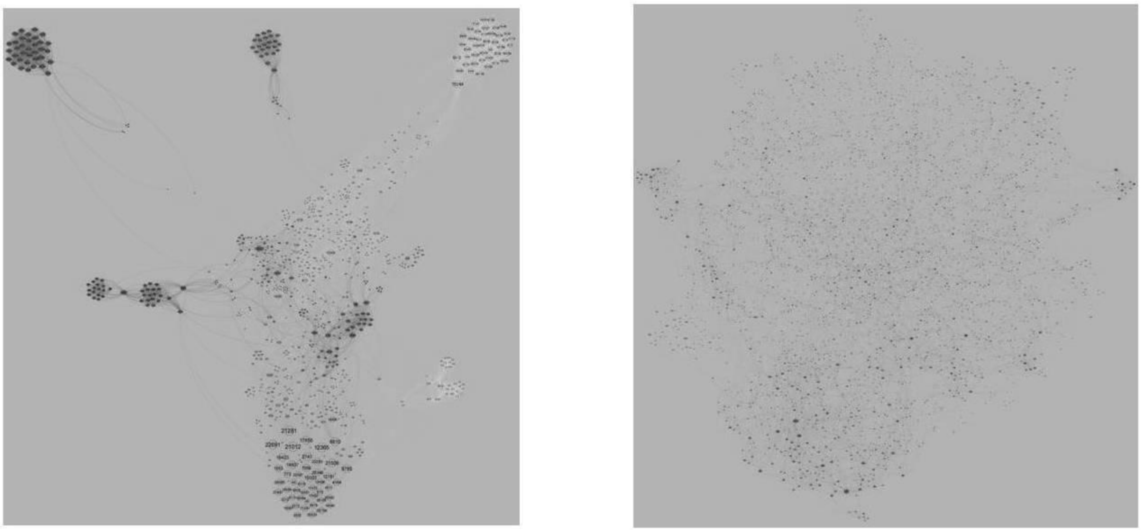
where $\| x \|$ indicates the sum of the elements of matrix x. This parameter measures the fraction of edges in the graph which connect vertices in the same community, minus the expected value of the same number of edges in the graph with the same community partitions but with random connections between their respective nodes.

If the number of intra-community edges shows no improvement on the expected value, then the modularity would be Q=0. On the other hand, Q approaches a maximum value of 1 when the community structure is strong. According to [88], the usual empirical range for Q is between 0.3 and 0.7. The modularity is calculated after each iteration of the elicitation algorithm, when two new components have been created due to the elimination of an edge. At this point a test is made to see if a global maximum, or some predefined expected maximum, has been reached.

Discussion: the major problem of this algorithm is the computational cost which grows as the potential of the number of edges in the network being analyzed. If the network has a clear community structure, it will be divided into components after the first iterations of the algorithm. However, for more disperse networks, the computation cost can become very significant. This problem can be partially mitigated by sampling and/or early termination of the algorithm, as was proposed by Martínez-Arqué and Nettleton in [92].

Example community structure extraction and visualization: In [92], Martínez-Arqué and Nettleton presented a benchmarking of Newman’s method, which used sampling as a pre-processing step. In the following, and with reference to Figs. 6a and 6b, we briefly summarize their results. For the ArXiv-GrQc dataset[23] and with reference to Fig. 6a, the optimum modularity using Newman’s method was found at iteration 56 with Q=0.777 (modularity value), which partitioned the sampled version of the dataset in 57 communities. With reference to Fig. 6a, the greatest community (lowest part of the Figure), included 16.29% of the nodes in the complete graph. The following two communities by number of nodes together represent 10% of the graph.

For the Facebook dataset, with reference to Fig. 6b the optimum modularity was found at iteration 40, with Q=0.87, resulting in a total of 190 communities. This value was influenced by the low clustering coefficient of the complete graph.



(a)

(b)

Fig. 6. Newman’s Method: (a) Visualization of the principal communities extracted from a sampled arXiv-GrQc citation dataset; (b) Visualization for a sampled Facebook writes to wall log [92]

```
--------------------------------------------------------------
    Community Search Algorithm (Newman[88])
--------------------------------------------------------------
Find_communities:
WHILE modularity $\geq$ 0 OR
          current_iteration < max_iterations DO:
    1.1 Select the component to be considered.
    Initially this will be the biggest component.
    1.2 Calculate the edge similarity of the
    component {Edge_betweenness_edge_erasing}
    until it becomes divided.
    1.3 Save the two new components and
    eliminate the one which was analyzed.
    1.4 Calculate the {Modularity_Q}.
EWHILE

    Once the iterative process has terminated, assign
    the nodes in the network to the communities.

Edge_betweenness_edge_erasing:
WHILE the analyzed component is not divided in
          two parts DO:
    2.1 Calculate the betweenness scores of the
    edges
    2.2 Find the edge $E_h$ with the highest score
    2.3 Eliminate $E_h$ from the current component
    2.4 Recalculate the betweenness of the edges
    and  return to step 2.2.
EWHILE

Modularity_Q:
Generate the matrix of communities with size $k^2$
where $k$ is the number of components of the
current network:
    3.1 Calculate the Trace of the matrix of
    communities created previously
    3.2 Calculate the sum of the matrix
    3.3 Calculate the modularity
```

Other results in the literature for community extraction using the same or similar datasets are as follows: for the arXiv-GrQc dataset, Xie[93] reported 499 communities using a “Label Propagation Algorithm” and 605 communities using a “Clique Propagation Algorithm”. For the Facebook dataset, Viswanath in [55] reported a high fragmentation into small communities, and Leskovec in [94] also reported a relatively high fragmentation of communities in other OSNs akin to Facebook. In [92], the authors commented that the fragmentation of this particular dataset was also probably influenced by the interaction metric (writes to wall) used to define the strength of activity between users.

2.4.10.2 The ‘Louvain’ method [89]: this can be considered an optimization of Newman’s method, in terms of computational cost. Firstly, it looks for smaller communities by optimizing modularity locally. As a second step, it aggregates nodes of the same community and builds a new network whose nodes are the communities. These two steps are repeated iteratively until the modularity value is maximized. The optimization is based on evaluating the modularity gain, which is done by performing a local calculation of the change in modularity for a given community, caused by moving each node from it to an adjacent community. With each iteration the number of nodes to test quickly reduces (due to the aggregation of the corresponding nodes), and the computational cost is reduced in the same order.

3. Hot topics and latest work

In this Section we consider leading edge research and hot topics, based on papers presented in a selection of the major congresses in the last two years (2010-2012), which include OSN analysis themes: communities; influence and recommendation; models, metrics and dynamics; behaviour and relationships; and information diffusion.

The ‘hot topic’ references are taken from the following congresses: WWW 2012, ICWSM 2012, WOSN 2010 and WCCI 2012.

3.1. Communities

As would be expected, much of the most recent research on OSN communities is dedicated to the data mining of datasets from the most popular current online applications, such as Twitter, Baidu, Epinions, BitTorrent and DBLP. Among the specific themes studied are community similarity metrics, clustering methods, models and detection algorithms. This research is potentiated by the ever growing availability of datasets for analysis, and commercial motivations.

The following selected is summarized: a study of online expert panels in question answering communities[95]; Cross-community influence in discussion forums[96]; Community detection in incomplete information networks[97]; the economics of BitTorrent communities[98]; using content and interactions for discovering communities in Twitter and the Enron Email corpus[99]; a scalable constraint-based clustering algorithm based on a global similarity measure that takes into consideration the users' constraints and their importance in social networks[100]; an analysis and comparison of two applications dedicated to 'community question answering' (CQA), 'Yahoo! Answers' and 'Baidu Zhidao'[101]; and finally, a spectral partitioning method to detect communities in a co-authorship network[102].

Firstly, Pal et al. [95] present a study of online expert panels in question answering communities, whose objective is to evaluate changes in the behavioural patterns of the users over time. Unsupervised machine learning methods are used to identify interesting evolution patterns which help to distinguish between experts. Supervised classification methods are used to show that the models based on how users evolve over time can be more effective at expert identification than the models that do not consider evolution.

Cross-community influence in discussion forums is the theme considered in [96]. The authors pose several questions: (i) how to identify communities which persistently affect other communities; (ii) given a specific community, which communities does it influence; (iii) which communities are dependent on the activity of others; and (iv) how to identify that a community is being increasingly influenced or incorporated into another community. Two measures are proposed: (a) the degree of community membership of the users; (b) the centrality of the users within each community.

Community detection in incomplete information networks is considered by Lin et al. in [97], who address the problem of how to learn a global metric which can be used to measure the distance between any pair of nodes in a given incomplete information network. This is done by solving an optimization problem, which makes use of auxiliary information obtained from the link relations in the network and the set of dissimilar node pairs.

By optimizing an objective function, a matrix M is obtained which serves as a metric which can be used to measure the distance between any two nodes in the graph. This idea is borrowed from density-based clustering approaches which cluster nodes from a higher to a lower density. The distance-based clustering approach, called DSHRINK, is designed to detect the overlapped and hierarchical communities subjacent in the graph.

In order to reduce the computational cost of the clustering process, an approximation is allowed in the determination of the mutual nearest neighbours and local communities.

The method is tested on two DBLP Datasets, which provide bibliographic information on computer science journals and proceedings.

From this data, the authors derive a new dataset with incomplete information, obtained by applying a Snowball sampling process to extract one group of connected local regions at a time. A node is randomly selected and a Breadth First Search is performed to incorporate neighbouring nodes into the sample until a fixed number of nodes are obtained. This process is repeated until a given number of local regions have been sampled. As a consequence, the links within the local regions are maintained whereas the remaining links in the network are lost.

For empirical testing, Lin et al. used k-Means with the Euclidean metric is compared against the proposed algorithm using (i) a diagonal Mahalanobis matrix and (ii) a full Mahalanobis matrix. The authors claimed that with these methods their results showed that it is possible to learn a metric (represented in the M matrix) which can be used as input to a clustering algorithm in order to elicit the communities, with only a small information loss with respect to the complete dataset.

The economics of BitTorrent communities, which are private file-sharing communities built on the BitTorrent protocol, is considered in [98] by Kash et al.. It is observed that some of these communities have developed their own policies for motivating members to share content and contribute resources, such as the requirement for members to maintain a minimum share ratio between uploads and downloads. In this way a private community establishes a 'credit system', which could be considered as a simple economic system. The authors state that previous studies of these communities have focused on the information technology aspects rather than the economic aspects. Hence they conducted an economic study of the DIME community which shares live concert recordings.

It was found that users take into consideration the cost differential when deciding which files to consume. The DIME community defined a significant difference between the cost of new and old files. Frequent visitors can quickly accumulate credit by consuming newer files, and users compensate the higher cost of older files by downloading more copies of newer files, and then preferentially consuming older files during specially designated periods.

From the study[98], Kash et al. proposed some community rule changes for the DIME community in order to improve its functionality. From an informatics point of view, the two main recommendations were (i) restricting access of new users to older files and (ii) increasing demand for files in general. From an economics point of view, the three key recommendations were (a) distribution of wealth through progressive taxation; (b) the providing of incentives to add new files to the system; and (c) the relaxation of some of DIME's `rigid' rules, for example, the number of torrents that are allowed to be seeded.

Sachan et al., in [99], considered using content and interactions for discovering communities in social networks. Two real datasets were processed, the first extracted from Twitter and the second being the Enron Email corpus.

A person’s membership to a community is conditioned by the social relationship, the type of interaction and the information communicated with other members of that community. The authors defined a generative model which facilitates the discovery of communities based on the discussed topics, interaction types and the social connections among people.

The authors defined four models: a topic user community model, two variants of a topic user recipient community model and a ‘Full TURCM” which represents a composite of the first three models.

The first model, ‘topic user community model’, performs a latent community discovery in a network using the content being discussed by users in the form of latent topics and the type of posts generated by them. From this, a topic extraction was performed and the role of each user was identified.

The first model was formally defined as follows: *W* is the set of words in the corpus; *X* is the set of interaction types observed on the social graph among the set *U* of users (senders); *Z* and *C* are the set of latent topic and community assignments for every post, respectively. Then the joint probability distribution of users, posts, interaction types, topics and community assignments is given by:

$$L = P(W, X, U, Z, C, \theta, \phi, \eta, \lambda|\alpha, \beta, \nu, \delta) \tag{8}$$

The last seven parameters of formula (8) were estimated by sampling the conditional distribution of each variable using a Block-Gibbs sampling based approximate inference. The topic assignment and community assignment were sampled for each post from a conditional distribution for the current assignment given the observation and other assignments. From this a Markov chain was defined in which the state transitions were simulated by repeatedly sampling from the conditional distributions.

Benchmarking was performed against two existing methods from the literature: CUT(Community-User-Topic) and CART (Community-Author-Recipient-Topic). The CUT model uses the semantic content of social graphs to discover communities, whereas the CART model combines both content and link information available in a social network.

In [100], Alsaleh et al. present a scalable constraint-based clustering algorithm based on a global similarity measure that takes into consideration the users’ constraints and the importance of these constraints in social networks. Each constraint’s importance was calculated based on the occurrence of the constraint in the dataset. Performance of the algorithm was demonstrated on a dataset obtained from an online dating website using internal and external evaluation measures. Their results claim that the proposed algorithm is able to increase the accuracy of matching users in social networks by 10% in comparison to other algorithms.

For a clustering solution C = $\{C_1 \ldots C_q\}$, taking into account the number of users in each cluster, a ‘closeness’ function was defined as:

$$\text{Closeness}(C)=\frac{\sum_{i=1}^{q}\frac{S(C_i)}{W(C_i)^r}\times|C_i|}{\sum_{i=1}^{q}|C_i|\times \text{ConsPur}_q} \tag{9}$$

where S(Ci) represents the occurrence of every distinct value in a cluster Ci, W(Ci) represents the number of distinct values in the cluster Ci, |Ci| means the number of users in cluster Ci, r is a positive real number called 'repulsion' to control intra-cluster similarity, *q* is the number of clusters and *ConsPur$_q$* is a measure used to find the purity of constraint in cluster q.

Also, the constraint purity measure for a cluster q was defined as follows:

$$\text{ConsPur}_q=1-\frac{X_1+\ldots+X_n}{n} \tag{10}$$

where X_n can be between 0 (where the constraint does not exist) and 1 (where the constraint exists with full weight) depending on the weighting of the constraint. As a result *ConsPurq* will get a value between 1 (where all constraints are satisfied) and 0 (where all constraints are not satisfied with full weight).

The algorithm was benchmarked against three methods from the literature: the k-means clustering algorithm, the Clope clustering algorithm, and a semi-supervised constrained clustering algorithm.

To summarize, the authors, Alsaleh et al., have developed a scalable weighted constraint-based clustering algorithm (WSCClust) to efficiently cluster users in social networks, which uses a global similarity measurement function and takes into consideration the users constraints in social networks. The method was tested on a dataset obtained from an online dating website using internal and external evaluation measures. The results showed that WSCClust performed better in clustering people compared to three existing clustering algorithms (k-means, Clope and SSCClust).

'Community question answering' (CQA) is the theme studied by Li et al. in [101], in which the authors performed an analysis and comparison of two applications, 'Yahoo! Answers' and 'Baidu Zhidao'. A comparison of the similarities and differences of the two communities was made, with respect to their influence on solving questions. Specific aspects were studied: (i) the social network structures of 'Yahoo! Answers' and 'Baidu Zhidao'; (ii) a comparison of the social community characteristics of top contributors; (iii) the identification of the behaviour of users in different categories in these two portals; and (iv) the identification of temporal trends. Also, the efficiency and effectiveness for answering questions was compared for 'Yahoo! Answers' and 'Baidu Zhidao'.

The social network structure of the data was first analysed in terms of the well known 'bow tie' structure representation. Then a centrality analysis was conducted to identify the top contributors (users who contribute a great number of answers or questions) and the composition and characteristics of those contributors was studied.

Following this first analysis, the Louvain method [89] was then applied to detect sub-communities (with a minimum number of 10 users) in the four principal categories of users identified. The corresponding categories in Yahoo! Answers and Baidu Zhidao are identified by a mapping process, and then a statistical analysis is performed on the four typical category pairs in each application to identify similarities and differences between their sub-communities.

The weighted entropy was calculated for all detected sub-communities as follows: Let S denote the number of sub-communities detected, SC_s represents the s^{th} sub-community and c_{ij} represents the number of questions users u_i have asked (answered, asked or answered) in category j. Then, the weighted mean entropy E_{wm} was defined as follows:

$$E_{wm}=\frac{|SC_S|}{\sum_{s=1}^{S}|SC_S|}\sum_{j=1}^{T}P_{sj}\ln(P_{sj}) \tag{11}$$

$$P_{sj}=\frac{C_{sj}}{|SC_S|} \tag{12}$$

$$C_{sj}=\sum_{u_i\in SC_S}\frac{c_{ij}}{\sum_{k=1}^{T}c_{ik}} \tag{13}$$

where T is the total number of categories. Also, the entropy was calculated for the largest sub-communities, SCΔ.

A *fitness* measurement is defined to represent the degree of matching for two categories, in which Fit(i, j) denotes the *Fitness* of mapping the j-th category of

Yahoo! Answers to the i-th category of Baidu Zhidao, thus:

$$\mathrm{Fit}(i,j)=\frac{2\cdot m(i,j)}{\sum_{i=1}^{X} m(i,j)+\sum_{j=1}^{Y} m(i,j)} \tag{14}$$

The following points summarize the work of Li et al[101]: (i) Sub-communities were identified in Yahoo! Answers and Baidu Zhidao based on a reduced number of categories; (ii) it was shown that Yahoo! Answers had a certain amount of asker-answerers while in Baidu Zhidao askers and answerers were dominant and there were much fewer asker-answerers; (iii) it was found that in Yahoo! Answers more people preferred to ask or answer questions in a greater diversity of categories while in Baidu Zhidao, most people (especially the top contributors) preferred to ask or answer questions in a few categories in which they were most interested. The results were derived from datasets dated 2008 to 2010 for both applications.

Krömer et al., in [102], employ a spectral partitioning method to detect communities in a co-authorship network. It was commented that spectral partitioning and clustering methods are often used for graph and matrix analysis, as well as to detect structures in real world networks and databases. The Spectral clustering method constructs graph partitions based on eigenvectors of the adjacency matrix. This makes use of the observation that bi-partitions of a graph are closely connected with the second eigenvector of the graph Laplacian, and hence this eigenvector can be used to partition a graph. However, it was found that the partitioning was highly dependent on the weighting of the underlying network.

The authors used an intuitive weighting scheme based on ‘ant colony optimization’ (ACO). They compared the communities found by spectral partitioning when using the ACO inspired weighting, with those found by using trivial weighting based on the number of interactions between the users (co-authors in the DLBP database).

Ant Colony Optimization (ACO) is a meta-heuristic method which has recently gained following in the computational intelligence community. It is based on certain behavioural patterns of foraging ants, in which the emulation of the ants’ behaviour can be used as a probabilistic computational technique for solving complex problems that can be reduced to finding optimal paths in graphs. Krömer et al. used the ACO inspired approach for assigning weights to edges for the co-authorship network. The ACO mechanisms of pheromone increase and pheromone evaporation were used to model how the strength of ties varied between authors over time.

A first algorithm (simple iterative spectral partitioning, SimpleISP) divides the initial connected graph into two sub-graphs, each containing vertices (and incident edges) with positive valuation and vertices (and incident edges) with negative valuation respectively. For benchmarking, the DBLP dataset was used (dated April 2010). It was stored in an XML format and pre-processed. All conferences were selected that were held by IEEE, ACM or Springer, giving a total of 9,768 conferences.

To illustrate the partitioning process, the communities were computed for two authors with distinct characteristics (one highly collaborative and highly active). They were selected on the basis of a previous analysis of the DBLP dataset. The communities to which the authors belonged were found to diverge during the final iterations of the SimpleISP algorithm when using the trivial weighting. The highly active author was found to belong to less well connected communities than the highly collaborative author. However, when using the ACO inspired weighting, the communities of both authors were found to be more similar in terms of connectivity.

Summary of research in community detection: in this Section we have seen that current researchers are defining their own community detection models, which are customized depending on the type of data being analyzed and the research objective. Pal et al. [95] use supervised classification methods, whereas Lin et al. [97] use a density clustering method to process incomplete information. Sachan et al. [99] define a model based on interactions between users, whereas Alsaleh et al. [100] perform clustering based on a customized similarity measure. Finally, Krömer et al. [102] employ a spectral partitioning method. To summarize, we can see that researchers are developing their own paradigms for defining, detecting and partitioning the data into communities, rather than using *per se* the community detection algorithms of Newman and Girvan [88] or Blondel et al. [89] we saw in the base topic Section 2.4.10.

3.2. Influence and recommendation

It is reasonable to assume that online social networks will reflect natural human social orders and characteristics, hence different researchers have analysed log data in order to identify manifestations of, for example, trust between individuals, mutual support and hierarchies (pecking orders).

The following selected research illustrates these themes: language effects and power differences in social interaction[103]; a general framework for estimating the prevalence of deception in online review communities[104]; and the use of recommendations to boost content spread in social networks[105].

The theme of language effects and power differences in social interaction is considered in [103], in which Danescu-Niculescu-Mizil et al. propose that these aspects are key to understanding social interaction within groups for online communities. The authors propose the hypothesis that the power of individuals in a network can be measured in terms of the degree to which their style influences that of others. The authors consider how conversational behaviour can reveal power relationships in two distinct domains: discussions among Wikipedians and arguments before the U. S. Supreme Court.

The two datasets processed are (i) discussions among Wikipedia editors, containing approx. 240,000 conversational exchanges; and (ii) oral arguments presented before the U.S. Supreme Court, containing 50,389 conversational exchanges among Justices and lawyers. In order to carry out an evaluation which is domain independent and which can be generalized, function word classes are used rather than domain-specific substantive content.

One initial hypothesis is that if person A communicates with person B and person B uses many articles, prepositions or personal pronouns, then person A will tend to increase his/her usage of these language constructs, even if person A does not consciously realize this.

In the case of the Wikipedia, a user can be promoted to administrator status through a public election, usually after extensive historical contribution to the community. By studying the communications of editors over time, the authors study how linguistic coordination behaviour changes when a user becomes an administrator.

In the case of the US Supreme Court, conversations are analyzed between judges and lawyers, taking into account gender and also whether it is perceived a priori that the judge will be favourable or not towards the lawyers case. This latter aspect is considered given that a judge often comes into a case with a general leaning toward one side or the other based on their judicial philosophy. The lawyers, through their preparation for the case will often be able to evaluate the leaning of the given judge.

One of the key findings was that in terms of gender, female lawyers tend to coordinate more than male lawyers when talking to Judges, whereas Judges tend to coordinate more towards male lawyers than towards female lawyers.

Finally, in the context of the Wikipedia, it was found that users coordinated more with a given user U after this user had been promoted to administrator status, than previously when user U was just a normal user.

In [104], Ott et al. present a general framework for estimating the prevalence of deception in online review communities. The framework is based on the output of a noisy deception classifier, trained using a SVM. Using this framework, the authors conducted an empirical study of the prevalence of deception among positive reviews in six popular online review communities (Orbitz, Priceline, Expedia, Hotels.com, Yelp and TripAdvisor), identifying factors which influence deceptive opinion.

A signal cost function was defined which employs an inference mechanism based on Bayes' Theorem and signal theory. The signal cost of positive online reviews was defined as a function of the posting costs and exposure benefits of the review community in which it is posted. Based on this theory, the authors defined two hypotheses, which were supported by the empirical results:

(i) Review communities with low signal costs (that is, low posting requirements and high exposure), such as TripAdvisor and Yelp, will have more deception than communities with high signal costs, such as Orbitz.

(ii) Increasing the signal cost will decrease the prevalence of deception. For example, by excluding reviews written by novice reviewers, the prevalence and the growth rate of deception is reduced in the corresponding community.

In [105], Ranu et al. consider the use of recommendations to boost content spread in social networks. The authors proposed that content sharing in social networks is an important mechanism for content propagation on the Internet.

However, the degree to which content is distributed throughout the network depends on the relationships and connectivity between nodes. The authors stated that many current schemes for recommending connections are based on the number of common neighbours and the similarity of user profiles, but without taking into account the volume of relevant content found by a given user.

Hence the authors proposed a novel algorithm for recommending connections that boosts content propagation in a social network while maintaining the relevance of the recommendations. A novel aspect of their approach is a search based on edges instead of nodes, with a bound on the number of incident edges per node.

The model is called RMPP (Restricted Maximum Probability Path), and was benchmarked against four existing methods: 'greedy', 'continuous greedy', 'degree based selection' and 'friend of friend'.

The following is a brief summary of each of these methods. In the case of the 'greedy' method, edges with the largest lift in the current set of edges in the graph, are added one at a time; in 'continuous greedy' (CG), edges are added to the original graph by considering δ intervals of width 1/δ, and in each iteration incrementing y_i values of edges e_i in a feasible edge set *Y* with the maximum sum of gradients $\sum_{e_i \in Y} \frac{\partial F}{\partial y_i}$.This is followed by randomized rounding; 'degree based selection' is based on adding edges between high degree node pairs, in which one of the members of the pairs possesses a given content c; finally, in 'friend of friend' based selection, node pairs are ranked by the number of common neighbours, then edges are added between unconnected node pairs ordered by the ranking.

Returning to the authors RMPP model, the 'content maximization problem' was defined as follows: given a graph G = (V, E) and a constant k, find an edge set $X \subseteq \{(i, j) : i, j \in V\}$ such that: (i) at most k edges from X are incident on any node in V, (ii) for each $(i, j) \in X$, $i \in N_j$ and $j \in N_i$ and (iii) f(X) is maximum.

In order to evaluate the performance of the method with respect to the different edge selection methods, a lift metric was define as:

$$\text{Lift}(X) = \frac{f(E \cup X) - f(E)}{f(E)} \times 100 \tag{15}$$

where *E* is the set of edges in the original graph *G*, *f(.)* is a content spread function and *X* is the set of recommendations computed by a given edge selection method.

For empirical testing, four OSNs datasets were used: Wikipedia, Flickr, Epinions, and Twitter. The authors stated that these datasets were chosen because they capture a variety of different social relations, such as trust, follower-following and friendship.

Summary of research in influence and recommendation: we recall that in the base research of Section 2.4.8, the focus was on the overall structure of Internet pages and emails, based on graph topological characteristics[83][84[24]. On the other hand, the hot topics have involved specific case studies which also take into account sociological and psychological aspects of how and why users interact, for example, in the domains of US Supreme court decisions and Wikipedia editors[103]. Another psychological aspect, that of deception, is considered in [104], in which Bayes Theorem and signal theory are employed to measure this factor in a reviewing community

Recommendation is of course a key aspect of facilitating the spread of content [105] through a social network, while maintaining the relevance of the content for the receiving users. This topic is also related to the Information Diffusion hot topic which we cover separately in Section 3.5.

3.3. Models, metrics and dynamics

Within this theme, some of the current key areas of interest are the use of geographical information, how networks form, and how to efficiently process large datasets. Some of the applications which are currently most popular for analysis are YouTube, LiveJournal and especially Twitter.

We briefly describe the following research: a model for OSNs based on the notion of embedding the nodes in a geometric space, using a link probability based on a ranking of the nodes[106]; a graph analysis based approach and metrics to study social networks taking into account geographic information[107]; a node distance estimation mechanism which maps nodes in high dimensional graphs to positions in low-dimension Euclidean coordinate spaces[108]; the analysis of a Twitter data log in order to study the effect of restrictions on the number of connections in OSNs[109]; the emergence of social conventions (‘retweet’ and ‘via’) in OSNs (Twitter)[110]; a study of the YouTube social network[111]; a study of the strategy behind how credit networks form[112]; a solution for predicting missing links for a social network in Social Media by using user activity data[113]; and the clustering of keywords in Tweets[114].

In [106], Bonato et al. present a model for OSNs based on the notion of embedding the nodes in a geometric space, and a link probability based on a ranking of the nodes. OSN users were defined as points in an m-dimensional Euclidean space. Each node has a region of influence, and nodes may be joined with a certain probability if they become positioned within each others region of influence. The nodes were ranked by their popularity from 1 to n, where 1 is the highest ranked node and n the number of nodes. Nodes that are ranked higher have larger regions of influence, and so are more likely to acquire links over time. Only undirected graphs were considered. The number of nodes n was fixed but the population was defined as dynamic: at each time-step, a node is created and one is removed.

Scellato et al., in [107], present a graph analysis based approach to study social networks with geographic information. They define some novel metrics to characterize how geographic distance affects social structure. The authors analyze four large-scale OSN datasets (BrightKite, FourSquare, LiveJournal and Twitter) and their results show that a significant percentage of users have short-distance links and that clusters of friends are often geographically close. The results show that location-based OSNs such as FourSquare and BrightKite tend to have more geographically confined triangles than social networks more focused on content production and sharing such as LiveJournal and Twitter. They propose some new metrics, “node locality” and “geographic clustering coefficient” which incorporate the geographic positioning and distances of the nodes.

Zhao and Zheng, in [108], propose a novel node distance estimation mechanism that effectively maps nodes in high dimensional graphs to positions in low-dimension Euclidean coordinate spaces. This allows the node distance computation to be performed in constant time. The work is motivated by the need to reduce the computational cost for shortest path calculations in large graphs.

Ghosh et al., in [109], study the effect of restrictions on the number of connections in OSNs by analyzing a Twitter data log. They use a network growth model based on preferential attachment in order to assess the effects of different types of restrictions on OSNs, which also supports the design of new restrictions of varying rigidity.

Kooti et al., in [110], study the emergence of the social conventions ‘retweet’ and ‘via’ in the Twitter OSN. Their key findings were: (i) retweeting conventions arise ‘organically’, as a consequence of the perceived need to forward other people’s tweets efficiently in Twitter; (ii) early adopters of the retweeting conventions were more active and well-connected than the remaining adopters or typical users; (iii) the majority of early and later adopters had a Twitter colleague who adopted the same convention prior to them, thus demonstrating that conventions mainly spread via internal social links in Twitter; (iv) the conventions spread through a dense network, thus avoiding “bottleneck” users.

In [111], Wattenhofer et al. study the YouTube social network, and find it to have distinct network characteristics with respect to traditional online social networks. Some examples of these distinct characteristics are homophily, reciprocative linking, and assortativity. However, Youtube was found to have similar characteristics to Twitter, the latter being another example of a content-driven online social network. By studying the social and content aspects of user popularity, they found a strong correlation between a user’s social popularity and his/her most popular content.

The strategy behind how credit networks form is considered by Dandekar et al. in [112]. The authors starting point was the modelling of abstract credit networks in order to capture the dynamics of trust and obligations among agents over a series of transactions. The key characteristics of a distributed credit model were cited as being: robustness to intrusion, bounded risk, and a wide distribution of transaction pairings including sparse direct credit relationships.

Dandekar et al. considered the question of how this type of networks come into being from the decisions of autonomous self-interested agents to grant credit to others, and where agents in the network transact only with other agents that they directly trust. Under this restriction the authors demonstrate that when the ‘Nash equilibrium’ is true, agents allocate credit budgets in a socially optimal manner. Another model is also considered, in which agents have common beliefs about the default risk of others. In this model it is found that in equilibrium, all parties tend to issue credit to the same agents.

Formally, a simple undirected graph G(V;E) is defined as representing the underlying graph, in which V contains n nodes (agents), and E contains m edges (relationships between agents). Each node is constrained to extend credit only to nodes it trusts, which are considered as being its neighbours in G, and in this way the total utility is maximized.

Activity in the network (giving credit to neighbouring nodes) converts G into a directed graph, G = (V;E), in which nodes are entities or agents and edges represent pair-wise credit limits between agents. An edge(u, v) $\in$ E has capacity c_{uv}> 0, which implies that u has extended a credit line of c_{uv} units to v in v’s currency. If a node y needs to pay p units in its currency to node z (for example, to buy a good that z is selling), the payment can go through if the maximum credit flow from z to y is at least p units. The payment will get routed through a chain of nodes from y to z, where each link on the chain carries at least the requisite credit capacity. If an edge (u, v) routes p units of payment from u to v, the credit capacity c_{uv} increases by p while c_{vu} decreases by p.

Dandekar et al. conducted a series of empirical game experiments in order to evaluate the behaviour of network formation. The first game experiment included the following eight possible strategies: (i) a ‘zero strategy’ in which an agent does not necessary have to extend any credit, but may participate in the credit network as long as other agents extend credit to it; (ii) a ‘random link strategy’ in which agents extend c = 5 units of credit to each of the other agents with probability ¼; (iii) a ‘low default strategy’ in which an agent extends 5 units of credit to other agents whose probability of defaulting is below a given threshold; (iv) a ‘high transaction probability’ strategy in which an agent extends 5 units of credit to other agents with whom the creditor has a high probability of transacting as a buyer; (v) a ‘high expected value’ strategy in which an agent extends 5 units of credit to the agents from whom the creditor expects to gain the most value; (vi) a ‘high transact and low default’ strategy which commences by eliminating all agents with a very high probability of defaulting and then issues 5 units of credit to the remaining agents with whom the creditor is most likely to transact; (vii) a ‘high expected value and low default’ strategy which eliminates the same high-default-risk agents from consideration, but then uses the ‘high expected value’ method to select which of the remaining agents to issue 5 units of credit; (viii) the ‘high expected value or low default strategy, which offers 5 units of credit to agents with especially low default probability and also to agents with especially high expected net transaction value.

To summarize the work of Dandekar et al., from the empirical experiments, it was found that players consistently chose the Low Default strategy in equilibrium. This resulted in a centralized credit network structure in which a few highly connected agents facilitate trade among the rest, in a similar manner to a central currency model. It was found that the central agents attracted credit allocations for two reasons: (i) they had relatively low default risk and (ii) as a consequence that they were receiving credit from many others in the network, this converted them into hub nodes with high connectivity to many other nodes in the network.

In [113], Kamei et al. propose a solution for predicting missing links for a social media network by using user activity data. The solution was based on a probabilistic model with latent features (traits) for simultaneously generating links and activities in the set of nodes. It employed an efficient method for learning the model from the observed links and activities. In order to estimate the total number of latent features and the probability distribution of them for each node from the observed data, a hierarchical Dirichlet process (HDP) was incorporated into the model. The learned model was then used to predict missing links in a social network. The experimental results used synthetic data and a Japanese word-of-mouth communication website for cosmetics (@cosme). The authors showed that the proposed learning method could accurately estimate the link creation probabilities when there was sufficient training data.

Miyamoto et al., in [114], consider the clustering of keywords in Tweets. A series of tweets was handled as a sequence of words and an inner product space was introduced to a set of keywords on the basis of positive definite kernels using a fuzzy neighbourhood defined on a given sequence. Two clustering methods were tested: agglomerative hierarchical clustering and c-Means. Pair wise constraints were introduced to improve the interpretability of clusters.

Summary of research in models, metrics & dynamics: in comparison with the base themes we saw in Sections 2.4.5 and 2.4.6, the most recent research is more domain specific and takes more into account the nature of the users as well as the topological graph characteristics and information propagation dynamics we saw previously in [58][59]. We see the incorporation of multimedia and geo-location information [107] and the study of the evolution of specific domains such as credit networks[112], also related to themes such as trust between users, which we considered in Section 3.2. Another research area is how to map a higher dimension space to a lower one[108], which is related to the theme of how to process large datasets ('big data') which we looked at in the base Section 2.4.7. Finally, we observe that Twitter is currently one of the most popular applications among researchers[109][110][114] for analysis purposes, partially due to the availability of datasets, APIs for extracting the data and 'off the shelf' analytical tools specifically for Twitter.

3.4. Behaviour and relationships

Three current aspects of interest of the theme of behaviour and relationships are 'location based analysis', identifying relations from how users interact, and how to form optimum teams from the human resources who are users of an OSN. Hence, in this subsection we briefly describe the following research: the analysis of historical data from location-based social networks (LSBNs)[115]; the prediction of relationships from social behaviour data[116]; and online team formation in social networks[117].

The analysis of historical data from location-based social networks (LSBNs) is considered in [115] by Gao et al.. LBSNs provide location related services that allow users to "check-in" at geographical locations and share their experiences with their friends. The authors defined two behaviour models: (i) a historical model (HM) and (ii) a social-historical model (SHM). The results were contrasted with three baseline models: (a) Most Frequent Check-in model (MFC), (b) Most Frequent Time model (MFT), and (c) Order-k Markov Model. The MFC baseline model considers the power-law property in terms of the rich-get-richer effect, the MFT model considers only the temporal pattern, whereas the Order-k Markov Model considers the short-term effect of historical check-ins.

Adali et al., in [116], deal with the prediction of relationships from social behaviour data, using a set of behavioural features that capture the "function" of a specific relationship without the need for textual features. The behavioural features were based on the statistical properties of communication patterns between individuals such as reciprocity, assortativity, attention and latency. A new methodology was presented for determining how such features could be compared to textual features, and a Twitter dataset was processed to illustrate how these features could be used to accurately capture the contextual information which is present in textual features.

In the context of Twitter, Adali et al. considered three principal actions: PAIR represents any exchange between the two individuals. It considers directed messages; CONV represents "conversations", which are defined as sustained exchanges of directed messages between two individuals in a short amount of time; (iii) PROP considers "propagations", that is, messages of any kind from A to B which are later propagated by B to a third party.

The following statistical behavioural features were defined. USER: user's network (USER-A, USER-B) measures social and behavioural features of the user which are not specific to a pair; A-ATTN: A's relative attention measures how much B is getting A's total attention; B-ATTN: B's relative attention measures how much of B's attention is given to messages from A; BAL: Balance measures the degree similarity of two users (assortativity); RECIP: Reciprocity measures to which extent a node reciprocates the actions of another; TIME: measures the actual time in hours it takes for a user to respond to another person; PRI: Priority measures to which degree a person prioritizes another person over all their acquaintances; DEL: Delay measures how much a user is typically delayed to get an answer or how many other messages are prioritized over a message from the given user.

Anagnostopolous et al., in [117], consider the problem of online team formation in social networks. The authors proposed solving this problem by formulating an algorithm that assembles teams of experts to deal with tasks. This is done in such as way so that the coordination costs are bounded and the workload is fairly allocated. In order to conceptualize the scenario, the authors formalize what they call a ‘balanced social task assignment’ problem and conduct empirical tests on two real-world datasets, IMDC and Bibsonomy. They claim that, compared with solutions that do not take into account user workloads, their algorithms achieve a significant decrease in the imbalance of the workload (60%–70%) while incurring only a small increase of the coordination overhead (5%–10%).

The algorithm works as follows: upon arrival of a new task J, a team is formed for task J by solving an instance of the ‘social task assignment problem’. The ‘social task assignment sub-problem’ for task J consists in selecting the team Q that minimizes a specific cost allocation function a(Q) subject to a constraint on the coordination cost c(Q). The social task assignment problem for task J is defined as follows:

$$\begin{gathered}\min_{Q} \mathrm{a(Q)} \\ \mathrm{cov}(J,q)=1 \\ c(Q)\leq B.\end{gathered} \tag{16}$$

Anagnostopolous et al. consider two models for online team formation. In the first model, denominated *“Implicitly Connected Teams (ICT)”,* each team Q is not required to form a connected graph. It is only necessary that the communication paths between people in the team are through other members of the social network (not necessarily present in Q). The model is considered realistic if the existence of a short chain of acquaintances is sufficient to declare people as “compatible.”

The second model, called *“Explicitly Connected Teams* (ECT)”, requires that each team Q forms a connected graph using the set of edges T ={ (p^i, p^j) : p^i, $p^j \in$ Q}, and the distances computed in sub-graph (Q, T) are designated by d.

In computational terms, the ICT model computes the coordination cost over the whole social network, whereas the ECT model only computes the coordination cost over the sub-graph induced by the team.

Two measures, designated Steiner(Q) and Diam(Q), were considered for the coordination cost of a given team Q.

1. Steiner(Q) is the cost of the min cost Steiner tree *T* whose terminal nodes are the team members.

2. Diam(Q) = $\max_{pi,pi'\in Q}$ d^j(i, i’) for which the objective is to find a team of users Q with diameter at most B which minimizes a(Q). In contrast to ‘Steiner’, ‘Diam’ does not build a new problem instance to combine the load with the social cost function - instead it solves the problem directly.

The empirical tests were performed on datasets extracted from the IMDB (Internet Movies) and Bibsionomy databases. A brief describe of how they are interpreted in the context of the paper follows.

In the case of IMDB, two types of movie personnel (experts) were defined as directors and actors, and the movie genres were used as the skills. For example, Alfred Hitchcock {director} has the skills of {comedy, crime, film-noir, mystery, romance, thriller}.

The Bibsonomy dataset, on the other hand, represents a social-bookmarking and publication-sharing system, and contains a large number of computer science related publications, each of which is written by a set of authors. The Bibsonomy website is visited by a large community of users who annotate the publications using tags, such as ‘theory’, ‘software’, or ‘ontology’. The set of tags associated with the papers of an author are used to represent the skill set for the corresponding author.

The empirical results presented by Anagnostopolous et al. plot the Steiner cost against max. load and average team size for the IMDB and Bibsonomy datasets, and the Diam cost Vs max. load.

Summary of research in behaviour and relationships: this theme is related to base Sections 2.4.6 (Modelling, evolution and structure), 2.4.8 (Influence and Recommendation) and hot topic 3.2 (Influence and Recommendation). We have seen in this Section that geographical location data is now ‘in vogue’ [115] given that this type of data has recently become available from smart phone applications. Where users visit and in what sequence can be represented by Markov type models[115].

Also, new types of interactions between individuals are considered [116], such as reciprocity, attention and latency for Twitter data. Finally, team formation is considered [117] for social tasks where the teams may be implicitly or explicitly connected.

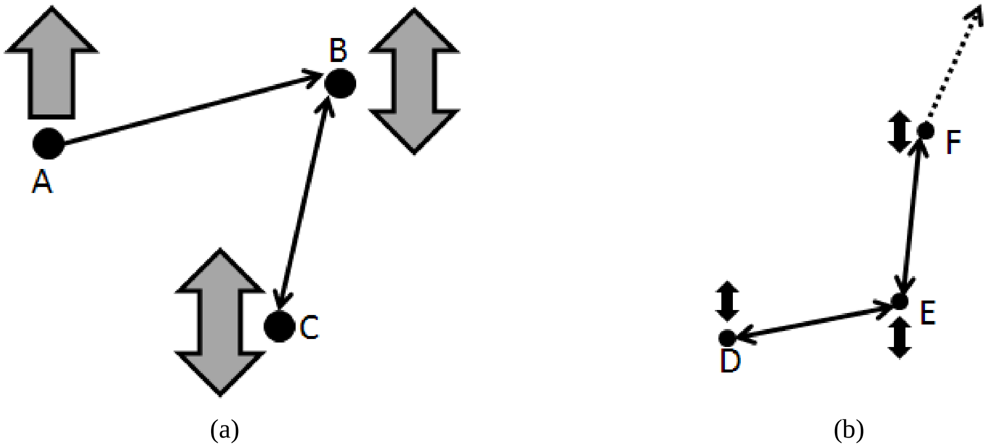


Fig. 7. Information Diffusion: (a) Spammer (A) has many outgoing links but very few incoming, whereas true hubs (B and C) have many outgoing and incoming links and intercommunicate between each other; (b) word of mouth viral communication may propagate via many low degree interlinked nodes.

3.5. Information diffusion

Information diffusion is a key aspect of interest for those who for commercial reasons wish to get the right publicity to the right people. Also it is of interest to analysts who want to track what is being transmitted through a given network and by whom.

We could say that on the one hand we have the ‘legitimate’ diffusion of information (solicited, such as news bulletins to which a user subscribes or relevant targeted commercial information) and on the other hand we have the ‘undesirable’ diffusion of information (unsolicited, such as SPAM, or ad-hoc badly targeted commercial information, criminal activity and under false pretences). Another focus of research is how users propagate information from one to another (analogous to ‘word of mouth’) and a promising way of analyzing this is based on ‘memes’.

In Fig. 7a we see how a high volume spammer can be easily identified in a network. One the other hand, Fig. 7b depicts how viral marketing uses bona fide users to spread a message by ‘word of mouth’. In a recent MIT study [118], advertisers are recommended to be social rather than commercial in their marketing messages. For example, advertising programs should use phrases such as ‘be like your friend’, ‘your friend knows this is a good cause’, ‘learn from your friend’ and ‘don’t be left out’, in order to provoke ‘viral’ propagation through the network.

In order to illustrate some of these themes, we briefly describe the following research: a model for representing information in OSNs[119]; the analysis of spammers’ social networks in order to identify criminal individuals and groups[120]; the role of social networks in information diffusion[121]; detecting ‘Internet buzzes’ or ‘amplification phenomena’[122]; contextual entity tracking using 'memes'[123], and its application to micro-blogging platforms such as Twitter[124]; finally, the categorization of Tweets in a large Twitter dataset[125].

A model for representing information in OSNs, which assigns two parameters to each information item, called endogeneity and exogeneity, is presented by Agrawal et al., in [119]. The endogeneity of an item quantifies its tendency to spread primarily through the connections between nodes. On the other hand, the exogeneity quantifies its tendency to be acquired by the nodes, independently of the underlying network. The authors extend the item-based model to take into account the openness of each node to new information. ‘openness’ is quantified by introducing the receptivity of a node as an additional parameter in the model. Given a social network and data related to the ordering of the adoption of information items by nodes, a maximum-likelihood based method is defined for estimating the endogeneity, exogeneity and receptivity parameters.

Yang et al., in [120], consider the analysis of spammers’ social networks in order to identify criminal individuals and groups. By analyzing inner social relationships in a criminal account community they find that criminal accounts tend to be socially connected, forming a small-world network. Also, criminal hubs, located in the centre of the social graph, are more inclined to follow criminal accounts. By analyzing outer social relationships between criminal accounts and their social friends outside the criminal account community, three categories of accounts are revealed which have close friendships with criminal accounts. From this a criminal account inference algorithm is defined by exploiting criminal accounts’ social relationships and semantic co-ordinations.

The “Mr.SPA” algorithm (Malicious Relevance Score Propagation Algorithm) propagates an MR score M_i to each node V_i after the initialization phase, using the following three score-assigning policies: Policy 1: MR Score Aggregation. An account’s score should sum up all the scores inherited from the accounts it follows. Policy 2: MR Score Dampening. The amount of MR score that an account inherits from other accounts should be multiplied by a dampening factor of α according to their social distances, where $0 < \alpha < 1$. Policy 3: MR Score Splitting. The amount of MR score that an account inherits from the accounts it follows should be multiplied by a relationship-closeness factor W_{ij}, which is the weight of the edge in the malicious relevance graph.

A second algorithm, called CIA, is designed to infer more criminal accounts based on a small known seed set, by analyzing the social relationships and semantic co-ordinations among accounts.

Bakshy et al., in [121], consider the role of social networks in information diffusion, conducting a large scale field experiment using 250 million Facebook users. The experiment randomizes exposure to signals about friend’s information sharing. A first finding is that those who are exposed are significantly more likely to spread information, and do so sooner than those who are not exposed.

A second finding is with respect to the role of strong and weak ties in information propagation. The authors confirm that stronger ties are individually more influential. However they find that is the weak ties, which are much more frequent, which account for the propagation of novel information. The users were demographically identified by gender, age and country.

The specific data analysis technique used was “temporal clustering”, which was used to identify the degree of proximity of the actions of the users.

Tie strength was measured in terms of four types of interactions: (i) frequency of private online communication between the two users in the form of Facebook messages; (ii) frequency of public online interaction in the form of comments left by one user on another user's posts; (iii) number of real-world coincidences captured on Facebook in terms of both users being labelled as appearing in the same photograph; and (iv) number of online coincidences in terms of both users responding to the same Facebook post with a comment. These four types of interactions are summarized as: comments received, messages received, photo coincidences and thread coincidences.

In [122], Lesot et al. address the task of detecting ‘Internet buzzes’, which are defined as ‘amplification phenomena’, that is, the diffusion on a very large scale of an Internet content, massively taken up within a short period of time. The authors propose two approaches based on the citation graph that represents hyperlinks relations between websites. The first approach detects temporal abnormalities in the number of citations of an information source, identifying information sources that undergo a surge of their direct citations. The second approach exploits higher level cues, based on the definition of the ‘dynamic cumulative visibility’ of an article, thus representing a ‘citation cascade’ which is characteristic of a ‘rumour’ type ‘buzz’. Both approaches are tested on real Web data (citation graph) and simulated data.

A ‘buzz’ is characterized by two key aspects: ‘intensity’ and ‘suddenness’. The citation graph is composed of nodes that correspond to Web pages representing information sources, and each source is composed of articles published by that source.

The method proposed by Lesot et al. is related to the area of contextual entity tracking, and its application to micro-blogging platforms such as Twitter. An information unit can be precisely identified (called a 'meme' [123]) and 'expressions' are employed to propagate ideas. There has recently been an increase in interest [124] in the concept of 'memes' and its use as a basis to analyze online social media interaction. A *meme* [123] is a concept which can be used to explain principles for the evolution and diffusion of ideas and cultural phenomena. Memes can be defined as recognisable cultural entities transmitted through imitation phenomena, via text, speech or gestures. It is said that the WWW has potentiated the creation and propagation of memes, in the form of textual word expressions, which has become an area of study with the objective of detecting and tracking 'hot topics'.

Twitter user and message logs are currently being analyzed by many authors as an OSN application However, Twitter messaging data tends to have a high volume, which includes noise and spam. Also, many Tweets are difficult to classify given the limited or abbreviated information which comprise the Tweets themselves. Poschko, in [125] processes a large Twitter dataset and classifies Tweets into specific categories: geo-location (Europe, Scandinavia, ...), person (Obama, Gates, ...), organization (Google, Greenpeace, ...), event (Easter, election, ...) and category (photography, politics, ...). The Python programming language was used to extract the Tweets from the raw data, a dataset collected over a 2 year time span, consisting of 85000 hash tags corresponding to 2.8million Tweets. Co-occurrence analysis was performed on the hash tags, and a machine learning approach (maximum entropy classifier) was used to classify them. A two step classification was performed: in Step 1 tags are classified in terms of Tweets and in Step 2 the tag is classified as the average of all classifications in Step 1.

Summary of research in information diffusion: this theme is related to the base Section 2.4.8 (Influence and Recommendation) and hot topic Section 3.2 (Influence and Recommendation). In the base Section 2.4.8, the focus was on the overall structure of Internet pages and emails, based on graph topological characteristics. Contrastingly, in the recent research we have seen in this Section, there appears to be a greater interest in the information theory properties of information items for propagation purposes. For example, properties such as 'endogeneity' and 'exogeneity' as defined in [119]. There is also an interest in how OSNs communities are related to information diffusion[121], viral type diffusion[122] and memes as transferable information units in [123][124].

4. Summary and concluding remarks

The analysis of OSN data is still in its infancy. However, it has a solid basis and starting point in disciplines such as graph theory and social psychology. There are also strong motivations, such as how to efficiently propagate the right information to the right people, and the consolidation of the use of OSNs by a growing percentage of the population, all of which foresee that it will become a research area of increasing importance.

On the one hand, algorithms such as VF2 for isomorphic matching and the Louvain method for community detection, make it possible to efficiently process the large data volumes which are found in the typical logs generated by OSN applications such as Twitter, Facebook, LinkedIn, Flickr, LiveJournal, Baidu, Epinions, BitTorrent and DLBP.

On the other hand, the dynamicity of OSNs make it an exciting field to work in, identifying real trends in individual and collective user behaviour and interactions, and how the OSN networks evolve over time.

The tendencies that we have identified from the research presented in the hot topic Sections include the analysis of specific user domains, and the study of socio-psychological aspects such as trust, deceit, hierarchies, and so on. Also the more topological and structural considerations of the graph have been superseded by the employment of information theory concepts to solve problems such as how to define and characterize information items in order to optimize their propagation.

We note that those who work in the research departments of large Internet companies such as Facebook, Twitter, Yahoo, Google, Microsoft, and so on, will have an advantage over those whose university computers have more limited processing capacity and who have less access to real OSN data logs. This can be mitigated by the availability of 'in the cloud' processing such as that offered by BigML [126] and APIs which facilitate the 'scraping' of some OSN applications (Twitter, LinkedIn).

We envisage that as OSNs become more and more a part of our everyday lives, they will capture more accurately the real interactions which are generated between groups of human beings and therefore will become a virtual mirror of reality which can be profitably analyzed.

However, on the other hand we could say OSNs will always lack the complete picture of what users are doing, as there are many forms of interaction, technological and otherwise, which are missing from the log datasets. Cross application studies and the addition of external demographic and dynamic geo-location data can help to mitigate this difficulty.

Acknowledgments

This research is partially supported by the Spanish MEC (project HIPERGRAPH TIN2009-14560-C03-01).

References

[1] D. Easley, J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010.

[2] T.h. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms (2nd ed.), MIT Press and McGraw–Hill, ISBN 0-262-53196-8, 2001.

[3] D.B. West, Introduction to Graph Theory - second edition, Prentice-Hall, Englewood Cliffs, NJ., 2001.

[4] M. Tomassini, Introduction to graphs and networks. Information Systems Department, HEC, University of Lausanne, Switzerland, 2010, available at https://www.hec.unil.ch/docs/files/6/662/3031_intro_graphs.pdf

[5] S. Wasserman, K. Faust, Social Network Analysis in the Social and Behavioral Sciences, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994, pp. 1–27.

[6] J. Farganis, Readings in Social Theory: the Classic Tradition to Post-Modernism, McGraw-Hill, New York, 1993.

[7] G. Simmel, Die Grosstädte und das Geistesleben (The Metropolis and Mental Life), Dresden: Petermann, 1903.

[8] J. Levy Moreno, Who shall survive?: Foundations of sociometry, group psychotherapy and sociodrama, Beacon House, 1953.

[9] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annual Review of Sociology, Vol. 27, 2001, pp. 415-444.

[10] G. Robins, P. Pattison, J. Woolcock, Small and Other Worlds: Global Network Structures from Local Processes. American Journal of Sociology (AJS), Volume 110, Number 4 (2005), 894-936.

[11] http://en.wikipedia.org/wiki/Social_networking_service

[12] M. Hauben, R. Hauben, T. Truscott, Netizens: On the History and Impact of Usenet and the Internet (Perspectives), Wiley-IEEE Computer Society P., ISBN 0-8186-7706-6, 1997.

[13] K. Hafner, The WELL: A Story of Love, Death and Real Life in the Seminal Online Community, Carroll & Graf Publishers, ISBN 0-7867-0846-8, 2001.

[14] O. El Akkad, So Long, GeoCities, The Globe and Mail, published Oct. 02 2009, available at: http://www.theglobeandmail.com/technology/globe-on-technology/so-long-geocities/article790790/

[15] D. M. Boyd, N.B. Ellison, Social network sites: Definition, history, and scholarship, Journal of Computer-Mediated Communication, 13(1), article 11, 2007, available at: http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html.

[16] E. Knapp, A Parent's Guide to Myspace?, DayDream Publishers, ISBN 1-4196-4146-8, 2006.

[17] R.E. Wilson, S.D. Gosling, L.T. Graham, A Review of Facebook Research in the Social Sciences, Perspectives on Psychological Science, May 2012, vol. 7, no. 3, pp. 203-220, available at: http://pps.sagepub.com/content/7/3/203.

[18] W.W. Zachary, W.W., An Information Flow Model for Conflict and Fission in Small Groups. Journal of Anthropological Research, Vol. 33 (1977) 452-473.

[19] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait? Behavioral Ecology and Sociobiology, Springer Volume 54, Number 4, (2003), pp. 396-405.

[20] M. Girvan, M.E.J. Newman, Community Structure in social and biological networks, In Proc. National Academy of Sciences of the USA (PNAS), Vol. 99, No. 12 (2002), 7821-7826.

[21] Stanford Network Analysis Platform Datasets. Available at http://snap.stanford.edu/data/index.html.

[22] J. Gehrke, P. Ginsparg, J. M. Kleinberg. Overview of the 2003 KDD Cup. SIGKDD Explorations 5(2): 149-151, 2003.

[23] J. Leskovec, J. Kleinberg, C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.

[24] J. Shetty, J. Adibi, Discovering Important Nodes through Graph Entropy- The Case of Enron Email Database, In Proc. 3rd Int. Workshop on Link Discovery (LinkKDD '05), 2005, pp. 74 - 81.

[25] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, J. Leskovec, Mobile call graphs: beyond power-law

and lognormal distributions, In Proc. 14th ACM SIGKDD (KDD ’08), 2008, pp. 596–604, New York, NY, USA.

[26] M. Richardson and R. Agrawal and P. Domingos. Trust Management for the Semantic Web. In Proc. 2nd Int. Semantic Web Conference,ISWC, 2003, pp.351-368.

[27] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, Group Formation in Large Social Networks: Membership, Growth, and Evolution. In Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), pp. 44-54, ACM New York, NY, USA.

[28] J. Leskovec, D. Huttenlocher, J. Kleinberg. Signed Networks in Social Media. In Proceedings of CHI 2010, SIGCHI Conference on Human Factors in Computing Systems, pp.1361-1370, ACM New York, NY, USA.

[29] J. McAuley, J. Leskovec. Image Labeling on a Network: Using Social-Network Metadata for Image Classification. In Proc. ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Part IV. Lecture Notes in Computer Science 7575 Springer 2012, pp. 828-841

[30] J. Yang, J. Leskovec. Temporal Variation in Online Media. In Proc. WSDM '11 4th ACM Int. Conf. on Web Search and Data Mining, pp. 177-186, ACM New York, NY, USA.

[31] Twitter development APIs.
Available at https://dev.twitter.com/.

[32] LinkedIn developer APIs.
Available at http://developer.linkedin.com/apis.

[33] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks. Proc. 3rd. Int. AAAI Conference on Weblogs and Social Media, 2009, pp. 361-362.

[34] NetMiner 4, software tool for exploratory analysis and visualization of network data.
Available at http://www.netminer.com.

[35] ‘Neo4J’ Graph Database System. Available at http://neo4j.org/.

[36] A. A. Hagberg, D. A. Schult, P. J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proc. 7th Python in Science Conference (SciPy2008), Pasadena, CA, USA, Aug. 2008, pp. 11–15.

[37] JUNG, Java Universal Network/Graph Framework.
Available at http://jung.sourceforge.net/.

[38] ‘igraph’ library and API.
Available at http://igraph.sourceforge.net/.

[39] Stanford Network Analysis Platform (SNAP). Available at http://snap.stanford.edu/snap/index.html .

[40] D.J. Cook, B.L. Holder, editors, Mining Graph Data, John Wiley & Sons, Hoboken, New Jersey, 2007.

[41] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, Journal of the American Society for Information Science and Technology, Vol. 58, Issue 7, (2007), pp. 1019–1031.

[42] C.V. Ramamoorthy, Analysis of Graphs by Connectivity Considerations, Journal of the ACM (JACM) , Volume 13, Issue 2, April 1966, pp. 211 - 222.

[43] H. Tong, S. Papadimitriou, J. Sun, P.S. Yu, C. Faloutsos, Colibri: fast mining of large static and dynamic graphs, in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’08), 2008, pp. 686–694, New York, NY, USA.

[44] V. Nair, D. Mahajan, S. Sellamanickam, A Unified Approach to Learning Task-Specific Bit Vector Representations for Fast Nearest Neighbour Search, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 929-938.

[45] D. Gibson, J.M. Kleinberg, P. Raghavan, Inferring Web Communities from Link Topology, In Proceedings of the ninth ACM conference on Hypertext and hypermedia (Hypertext '98): links, objects, time and space, 1998, pp. 225 - 234.

[46] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A.S. Tomkins, The web as a graph: measurements, models, and methods, In Proceedings of the 5th Annual International Conference on Computing and Combinatorics (COCOON'99), 1999, pp. 1-17.

[47] D. Chakrabarti, C. Faloutsos, Graph mining: Laws, generators, and algorithms, ACM Computing Surveys, 38(1), Article 1, March 2006.

[48] X. Yan, J. Han, gSpan: Graph-Based Substructure Pattern Mining, in Proc. Second IEEE International Conference on Data Mining (ICDM'02), 2002, pp. 721.

[49] A. Inokuchi, T. Washio, H. Motoda, An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, Lecture Notes in Computer Science, 2000, Volume 1910/2000, pp. 13-23.

[50] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and Analysis of Online Social Networks, in Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC '07), 2007, pp. 29 - 42.

[51] D. Eppstein, J. Wang, J., Fast Approximation of Centrality. Journal of Graph Algorithms and Applications, Vol. 8, Nº 1, (2004), pp. 39-45.

[52] M.E.J. Newman, J. Park, Why social networks are different from other types of networks. Phys. Rev. E68, 036122, 2003.

[53] R.I.M. Dunbar, Coevolution of neocortical size, group size and language in humans. Behavioral and Brain Sciences, 16,4 (1993) 681–735.

[54] D. Chakrabarti, Y. Zhan, C. Faloutsos, R-mat: A recursive model for graph mining, in Proc. SIAM Data Mining Conference, 2004. SIAM, Philadelphia, PA.

[55] B. Viswanath, A. Mislove, M. Cha, K.P. Gummadi, On the Evolution of User Interaction in Facebook, in Proceedings of the 2nd ACM workshop on Online Social Networks WOSN’09, Barcelona, Spain, 2009, pp. 37-42.

[56] G. Kossinets, D. Watts, Empirical analysis of an evolving social network, Science, Vol. 311, No. 5757, (2006), pp. 88-90.

[57] L. Tang, H. Liu, J. Zhang, N. Nazeri, Community evolution in dynamic multi-mode networks, in Proc. of the 14th ACM SIGKDD (KDD ’08), 2008, New York, NY, USA., pp. 677–685.

[58] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, in Proc. KDD '05, 11th ACM SIGKDD Int.

Conf. of Knowledge Discovery and Data Mining, 2005, pp. 177-187.

[59] M. McGlohon, L. Akoglu, C. Faloutsos, Weighted Graphs and Disconnected Components: Patterns and a generator, in Proc. 14th ACM SIGKDD Int. Conf. of Knowledge Discovery and Data Mining (KDD '08), 2008, pp. 524-532.

[60] R. Kumar, J. Novak, A. Tomkins, Structure and Evolution of Online Social Networks, Link Mining: Models, Algorithms, and Applications 2010 (Springer), Part 4, pp. 337-357.

[61] M. Randic, L. M deAlba, L., Dense Graphs and Sparse Matrices, J. Chem. Inf. Comput. Sci. 37, (1997) 1078-1081.

[62] M.E.J. Newman, The structure and function of complex networks, SIAM Review 45, (2003), pp. 167–256.

[63] W. Hwang, T. Kim, M. Ramanathan, A. Zhang, Bridging centrality: graph mining from element level to group level, in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08), 2008, pp. 336–344, New York, NY, USA.

[64] Y-Y Ahn, S. Han, H. Kwak, Y-H Eom, S. Moon, H. Jeong, Analysis of Topological Characteristics of Huge Online Social Networking Services, WWW '07 (2007), In Proc. 16th International Conference on World Wide Web, pp. 835 - 844.

[65] B. Goncalves, N. Perra, A. Vespignani, Validation of Dunbar's number in Twitter conversations. arXiv.org-physics-arXiv:1105.5170, 28 May 2011 (http://arxiv.org/abs/1105.5170).

[66] J. M. Kleinberg, The Small-World Phenomenon: An Algorithmic Perspective. In Proceedings of the thirty-second annual ACM symposium on Theory of computing (STOC '00), 2000, pp. 163 - 170.

[67] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Weiner, Graph Structure in the Web, Computer Networks - International Journal of Computer and Telecommunications Networking, 33(1-6), pp. 309-320, 2000.

[68] J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, J. Zhanga, On graph problems in a semi-streaming model, Theoretical Computer Science, Volume 348, Issues 2-3, (2005), pp. 207-216.

[69] C. Demetrescu, I. Finocchi, A. Ribichini, Trading off space for passes in graph streaming problems, in ACM-SIAM Symposium on Discrete Algorithms, 2006, pp. 714–723.

[70] A. Das Sarma, S. Gollapudi, R. Panigrahy, Estimating PageRank on graph streams, in ACM Symposium on Principles of Database Systems, 2008, pp. 69–78.

[71] L.A. Goodman, Snowball sampling, Annals of Mathematical Statistics 32 (1): 148–170, 1961.

[72] T.A.B. Snijders, Estimation on the Basis of Snowball Samples: how to weight? Bulletin de Méthodologie Sociologique, Nº 36, 1992, pp. 59-70.

[73] T. Shafie, Design-based Estimators for Snowball Sampling, Workshop on Survey Sampling Theory and Methodology, Vilnius, Lithuania, 2010, August 23-27, available at http://vilniusworkshop2010.stat.gov.lt/Straipsniai/Shafie_T.pdf.

[74] K. Bartz, J. Blitzstein, J. Liu, Graphs, Bridges and Snowballs: Monte Carlo Maximum Likelihood for Exponential Random Graph Models, 2009, presentation, available at http://www.kevinbartz.com/uploads/graph/presentation.pdf.

[75] T.A.B. Snijders, Conditional Marginalization for Exponential Random Graph Models, Journal of Mathematical Sociology, 34, (2010) 239-252.

[76] E.W. Dijkstra, A note on two problems in connexion with graphs. Numerische Mathematik, Volume 1, Number 1, (1959) 269-271.

[77] F. Lu , P-C. Lai, P-C, A Shortest Path Searching Method with Area Limitation Heuristics, Lecture Notes in Computer Science, Volume 3991/2006, 884-887.

[78] L.S. Buriol, G. Frahling, S. Leonardi, A. Marchetti-Spaccamela, C. Sohler, Counting triangles in data streams, in ACM Symposium on Principles of Database Systems, 2006, pp. 253–262.

[79] H. Tong, C. Faloutsos, Center-piece sub graphs: problem definition and fast solutions, in Proceedings of the Twelfth ACM SIGKDD International Conference on KDDM, 2006, pp. 404-413.

[80] L.P. Cordella, P. Foggia, C. Sansone, M. Vento, Evaluating Performance of the VF Graph Matching Algorithm, Proc. of the 10th International Conference on Image Analysis and Processing, IEEE Computer Society Press, 1999, pp. 1172-1177.

[81] L. P. Cordella , P. Foggia C. Sansone, M.Vento, An Improved Algorithm for Matching Large Graphs, in Proc. 3rd IAPR-TC-15 International Workshop on Graph based Representations, Cuen, Italy, 2001, pp. 149-159.

[82] T. Washio, H. Motoda, State of the art of graph-based data mining, ACM SIGKDD Explorations Newsletter, Volume 5, Issue 1, 2003, pp. 59 - 68.

[83] J.M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Journal of the ACM (JACM), Volume 46, Issue 5, 1999, pp. 604 - 632.

[84] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03), 2003, pp. 137 - 146.

[85] J. Körner, Bounds and Information Theory, SIAM Journal on Algorithms and Discrete Mathematics, (7):560–570, 1986.

[86] P. Jaccard, Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines, Bulletin de la Société Vaudoise des Sciences Naturelles 37, (1901), pp. 241-272.

[87] L.C. Freeman, A Set of Measures of Centrality Based on Betweenness, Sociometry, Vol. 40, No. 1, (1977), pp. 35-41.

[88] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004.

[89] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebure, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment (10), 2008, pp. 1000.

[90] M.E.J. Newman, Modularity and community structure in networks, PNAS June 6, vol. 103 no. 23, (2006), pp. 8577-8582.

[91] J.M. Kleinberg, Challenges in mining social network data: processes, privacy, and paradoxes, Proc. 13th Int. Conf. on Knowledge Discovery and Data Mining (KDD '07), (2007), pp. 4 - 5.

[92] N. Martínez Arqué, D. F. Nettleton, Analysis of On-line Social Networks Represented as Graphs – Extraction of an Approximation of Community Structure Using Sampling, in Proc. Congress Modeling Decisions for Artificial Intelligence, MDAI 2012, LNAI 7647, pp.149-160, Springer-Verlag.

[93] Xie, J.,Szymanski, B.K. and Liu, X. 2011. SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker listener Interaction Dynamic Process. Cornell University Library http://arxiv.org, arXiv:1109.5720v3.

[94] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters, Internet Mathematics Vol. 6, No. 1 (2009) 29–123.

[95] A. Pal, S. Chang, J.A. Konstan, Evolution of Experts in Question Answering Communities, In: Proc. 6th Int. AAAI Conf. on Weblogs and Social Media, Dublin, Ireland, 4-7 June, 2012, pp. 274-281.

[96] V. Belák, S. Lam, C. Hayes, Cross-Community Influence in Discussion Fora, In: Proc. 6th Int. AAAI Conf. on Weblogs and Social Media, Dublin, Ireland, 4-7 June, 2012, pp. 34-41.

[97] W. Lin, X. Kong, P. Yu, Q. Wu, Y. Jia, C. Li, Community Detection in Incomplete Information Networks, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 341-349.

[98] I. Kash, J. Lai, H. Zhang, A. Zohar, Economics of BitTorrent Communities, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 221-230.

[99] M. Sachan, D. Contractor, T. Faruquie, L. V. Subramaniam, Using Content and Interactions for Discovering Communities in Social Networks, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 331-340.

[100] S. Alsaleh, R. Nayak, Y. Xu, Grouping People in Social Networks Using a Weighted Multi-Constraints Clustering Method, WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia, In: Proc. Int Conf. on Fuzzy Systems (FUZZ IEEE 2012), pp. 243-250.

[101] B. Li, M.R. Lyu, I. King, Communities of Yahoo! Answers and Baidu Zhidao: Complementing or Competing?, WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia, In: Proc. Int. Joint Conf. on Neural Networks (IJCNN 2012), pp. 524-531.

[102] P. Krömer, V. Snásel, J. Platos, M. Kudelka, Z. Horák, An ACO Inspired Weighting Approach for the Spectral Partitioning of Co-authorship Networks, WCCI 2012 IEEE World Congress on Computational Intelligence, June, 10-15, 2012 - Brisbane, Australia, In: Proc. Congress on Evolutionary Computation (IEEE CEC 2012), pp. 2477-2483.

[103] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, J. Kleinberg, Echoes of power: How power differences between people are revealed by linguistic style coordination, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 699-708.

[104] M. Ott, C. Cardie, J. Hancock, Estimating the Prevalence of Deception in Online Review Communities, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 201-210.

[105] S. Ranu, V. Chaoji, R. Rastogi, R. Bhatt, Recommendations to Boost Content Spread in Social Networks, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 530-538.

[106] A. Bonato, J. Janssen, P. Pralat, A Geometric Model for On-line Social Networks, In: Proc. 3rd Workshop on Online Social Networks (WOSN 2010), Boston, MA, USA (June 2010), http://www.usenix.org/events/wosn10/tech/full_papers/Bonato.pdf

[107] S. Scellato, C. Mascolo, M. Musolesi, V. Latora, Distance Matters: Geo-social Metrics for Online Social Networks, In: Proc. 3rd Workshop on Online Social Networks (WOSN 2010), Boston, MA, USA (June 2010), http://www.usenix.org/events/wosn10/tech/full_papers/Scellato.pdf

[108] X. Zhao, H. Zheng, Orion: Shortest Path Estimation for Large Social Graphs, In: Proc. 3rd Workshop on Online Social Networks (WOSN 2010), Boston, MA, USA (June 2010), http://www.usenix.org/events/wosn10/tech/full_papers/zhao.pdf

[109] S. Ghosh, G. Korlam, N. Ganguly, The Effects of Restrictions on Number of Connections in OSNs: A Case-Study on Twitter, In: Proc. 3rd Workshop on Online Social Networks (WOSN 2010), Boston, MA, USA (June 2010), http://www.usenix.org/events/wosn10/tech/full_papers/Ghosh.pdf

[110] F. Kooti, H. Yang, M. Cha, K.P. Gummadi, W.A. Mason, The Emergence of Conventions in Online Social Networks, In: Proc. 6th Int. AAAI Conf. on Weblogs and Social Media, Dublin, Ireland, 4-7 June, 2012, pp. 194-201.

[111] M. Wattenhofer, R. Wattenhofer, Z. Zhu, The YouTube Social Network, In: Proc. 6th Int. AAAI Conf. on Weblogs and Social Media, Dublin, Ireland, 4-7 June, 2012, pp. 354-361.

[112] P. Dandekar, B. Wiedenbeck, A. Goel, M. Wellman, Strategic Formation of Credit Networks, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 559-568.

[113] T. Kamei, K. Ono, M. Kumano, M. Kimura, Predicting Missing Links in Social Networks with Hierarchical Dirichlet Processes, WCCI 2012 IEEE World Congress on Computational Intelligence, June, 10-15, 2012 - Brisbane, Australia, In: Proc. Int. Joint Conf. on Neural Networks (IJCNN 2012), pp. 1816-1823.

[114] S. Miyamoto, S. Suzuki, S. Takumi, Clustering in Tweets Using a Fuzzy Neighborhood Model, WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia, In: Proc. Int Conf. on Fuzzy Systems (FUZZ IEEE 2012), pp. 251-256.

[115] H. Gao, J. Tang, H. Liu, Exploring Social-Historical Ties on Location-Based Social Networks, In: Proc. 6th Int. AAAI

Conf. on Weblogs and Social Media, Dublin, Ireland, 4-7 June, 2012, pp. 114-121.

[116] S. Adali, F. Sisenda, M. Magdon-Ismail, Actions speak as loud as words: Predicting relationships from social behavior data, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 689-698.

[117] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, S. Leonardi, Online Team Formation in Social Networks, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 839-848.

[118] C. Dunn, MIT: Facebook Still Has Large, Untapped Opportunity In Social Advertising, Content Marketing Institute, February 3, 2012, available at: http://www.contentmarketinginstitute.com/2012/02/facebook-social-advertising/

[119] R. Agrawal, M. Potamias, E. Terzi, Learning the Nature of Information in Social Networks, In: Proc. 6th Int. AAAI Conf. on Weblogs and Social Media, Dublin, Ireland, 4-7 June, 2012, pp. 2-9.

[120] C. Yang, R. Harkreader, J. Zhang, S. Shin, G. Gu, Analyzing Spammers' Social Networks For Fun and Profit - In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 71-80.

[121] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The Role of Social Networks in Information Diffusion, In: Proc. World Wide Web 2012 (WWW 2012), April 16-20, 2012, Lyon, France, pp. 519-528.

[122] M-J. Lesot, F. Nely T. Delavalladey, P. Capety, B. Bouchon-Meunier, Two Methods for Internet Buzz Detection Exploiting the Citation Graph, WCCI 2012 IEEE World Congress on Computational Intelligence, June, 10-15, 2012 - Brisbane, Australia, In: Proc. Int Conf. on Fuzzy Systems (FUZZ IEEE 2012), pp. 1368-1375.

[123] R. Dawkins, The Selfish Gene. Oxford University Press, 1989.

[124] J. Leskovec, L. Backstrom, and J. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” in Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2009, pp. 497–506.

[125] J. Poschko, Exploring Twitter Hashtags, 2010, available at http://twex.poeschko.com/media/files/ExploringTwitterHashtags.pdf .

[126] BigML, Machine Learning for Everyone, available at https://bigml.com/