Chapter 1

THE COMBINATION OF GRAPH THEORY AND UNSUPERVISED LEARNING APPLIED TO SOCIAL DATA MINING

Héctor D. Menéndez* ***and José Luis Llorente***
Universidad Autónoma de Madrid
Escuela Politécnica Superior. C/ Tomás y Valiente 11, 28049, Cantoblanco, Madrid

Abstract

Over the last few years, Social Data Mining has become an important field inside Data Analysis. These techniques are rapidly finding applications in a variety of domains including artificial intelligence, economics and marketing amongst others. They are based on knowledge extraction from the users, focusing on their behaviour and relationships inside a system which can be modeled as a Social Network where they act independly or establishing relationships. The Social Network studies have been oriented from different points of view, however, the most representatives come from Graph Theory. On the one hand, the Network is usually represented as a Graph where the users are considered nodes and their relationships are the graph edges. Different approximations of Complex Network Analysis are used to described the Network and its features. On the other hand, the Graph Theory can also be used to analyses the behaviour of the users, not only from a relationships point of view, instead, it can be used to analyse the information that they generates, creating an independent profile of the user. A representative selection of these techniques is discussed in detail in this work, showing how different methods extracted from Graph Theory can be combined with different approaches of Unsupervised Learning to analyse Social Behaviour from different perspectives.

1. Introduction

Social Data Mining is one of the most innovative areas in Data Mining [28]. This new field combines different techniques which are related to Data Mining and Complex Networks [42]. Several of these approaches have been focused on graph models,

*Corresponding Author. E-mail address: hector.menendez@uam.es

specially unsupervised learning, where graph models have proved to improve the results of the most classical algorithm [53].

Different steps of the Data Mining processes can be focused from a graph-based perspective. For example, the data structure can be consider as a Manifold generating a topology over the data instances considering them as nodes of a graph and the edges similarity measures amongst the nodes. New Data Mining algorithms from unsupervised learning based on Spectral Analysis [58] takes a graph as the topology of the data distribution and uses its spectrum for the model generation process of the Machine Learning algorithm.

Complex Network are usually used to represent a Social Network [43] (for example, Facebook or Twitter) as a graph where the users are represented as the nodes of the graph and the relations between the users are represented as the edges of the graph. This representation provides a lot of information about the network related to the type of network (Small-World [60], Random [19], Scale-free [5],...), and its features (strength, paths, authorities, hubs,...). It also has several applications related to really separated fields such as Marketing [54] and Medicine [56], amongst others.

This work is focused on give a general perspective about the influence of Graph Theory in Data Mining algorithms, presenting new techniques and algorithms which have been developed over the last few years. It also introduces some analytical methods extracted from Complex Networks and provides some examples of important algorithms and network structures in this field. Finally, it shows two famous Social Networks where some of theses techniques have been successfully applied (Facebook and Twitter) and recommends some software tools which might be helpful to apply these processes.

The chapter is structure as follows: Section 2 provides some basic definitions of Graph Theory which are useful to understand the rest of the paper, people with knowledge about Graph and Complex Network Theory can missed this section. Section 3 gives an overview of the Data Mining process and the different steps used during the analysis. This section introduces the steps of Data Mining. Next, the Machine Learning methods which can use Graph Theory are detailed in Sections 4. Section 5 is focused on the Complex Network analysis techniques which are also important in the Social Analysis, it explains some issues about the Network structure and presents two important algorithms of the literature: PageRank and HITS. Section 6 shows some practical application of these approaches in real-world Social Networks: Facebook and Twitter. Section 7 summarizes some tools which can be used in all the analysis processes. Finally, the last section, explains the conclusions.

2. Basic Definitions from Graph Theory

Some algorithms use concepts and metrics extracted from graph theory. For this reason, and before describing them, some of those basic concepts are briefly introduced.

Definition 2..1 (Graph)**.** A graph $G = (V, E)$ is a set of vertices or nodes V denoted by $\{v_1, \ldots, v_n\}$ and a set of edges E where each edge is denoted by e_{ij} if there is a connection between the vertices v_i and v_j.

Graphs can be directed or undirected. If all edges satisfy the equality $\forall i, j$,

$e_{ij} = e_{ji}$, the graph is said to be undirected.

The graph can also be represented through its adjacency matrix (the most usual approach) which can be defined as:

Definition 2..2 (Adjacency Matrix)**.** An adjacency matrix of G, A_G, is a square $n \times n$ matrix where each coefficient satisfies:

$$(a_{ij}) = \begin{cases} 1, & \text{if } e_{ij} \in E \\ 0, & \text{otherwise} \end{cases}$$

When it is necessary to work with weighted edges, a new kind of graph needs to be defined:

Definition 2..3 (Weighted Graph)**.** G is a weighted graph if there is a function $w : E \to R$ which assigns a real value to each edge.

Any algorithm that works with the vertices of a graph needs to analyse each node neighbours. The neighbourhood of a node is defined as follows:

Definition 2..4 (Neighbourhood)**.** If the edge $e_{ij} \in E$ and $e_{ji} \in E$ we say that v_j is a neighbour of v_i. The neighbourhood of v_i Γ_{v_i} is defined as $\Gamma_{v_i} = \{v_j \mid e_{ij} \in E \text{ and } e_{ji} \in E\}$. Then, the number of neighbours of a vertex v_i is $k_i = |\Gamma_{v_i}|$

Also nodes can generate paths between them through their edges, a path is defined as follows:

Definition 2..5 (Path)**.** A Path of a graph between the nodes v_i and v_j is a set of edges which connects these two nodes. It will be denoted by P_{ij}.

And its length is:

Definition 2..6 (Path Length)**.** The Path Length is defined as the number of edges contained in the path. It will be denoted by $|P_{ij}|$.

It is also important to know the shortest path between two nodes, usually defined by:

Definition 2..7 (Shortest Path)**.** The Shortest Path is a minimum Path between two nodes. It should satisfy:

$$\min_{|P_{ij}|}\{P_{ij}|\ P_{ij} \in G\} \tag{1}$$

One is most important metrics of the graph is defined by its diameter:

Definition 2..8 (Graph Diameter)**.** The Graph Diameter is defined as the maximum shortest path of the graph.

Once the most general and simple concepts from graph theory are defined, we can proceed with the definition of some basic measures related to any node in a graph: the average path length the clustering coefficient and the weighted clustering coefficient.

Definition 2..9 (Average Path Length)**.** Let G be a Graph and V its set of vertices. Let $d(v_i, v_j)$ be the shortest distance between v_i and v_j. The Average Path Length is defined by:

$$l_G = \frac{1}{n \cdot (n-1)} \cdot \sum_{i,j} d(v_i, v_j) \tag{2}$$

Definition 2..10 (Local CC [14])**.** Let $G = (V, E)$ be a graph where E is the set of edges and V the set of vertices and A its adjacency matrix with elements a_{ij}. Let Γ_{v_i} be the neighbourhood of the vertex v_i. If k_i is considered as the number of neighbours of a vertex, we can define the clustering coefficient ***(CC)*** of a vertex as follows:

$$C_i = \frac{1}{k_i(k_i-1)} \sum_{j,h} a_{jh} a_{ij} a_{ih} a_{ji} a_{hi} \tag{3}$$

The Local CC measure provides values ranging from 1 to 0. Where 0 means that the node and its neighbours do not have clustering features, so they do not share connections between them. Therefore, value 1 means that they are completely connected. This definition of CC can be extended to weighted graphs as follows:

Definition 2..11 (Local Weighted CC [6])**.** Following the same assumption of Local Clustering Coefficient definition, let W be the weight matrix with coefficients w_{ij} and A be the adjacency matrix with coefficients a_{ij}, if we define:

$$S_i = \sum_{j=1}^{|V|} a_{ij} a_{ji} w_{ij} \tag{4}$$

Then, the Local Weighted Clustering Coefficient can be defined as:

$$C_i^w = \frac{1}{S_i(k_i-1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{jh} a_{ij} a_{ih} a_{ji} a_{hi} \tag{5}$$

For this new definition, we are considering the connections between the neighbours of a particular node, but now we add information about the weights related to the original node. This new measure calculates the distribution of the weights of the node that we are analysing, and shows how good the connections of that cluster are. The following theorem proves that the weighted CC has the same value than the CC when all the weights are set to the same value:

Theorem 2..1. *Let G be a graph, A its adjacency matrix and W its weight matrix. If we set $w_{ij} = \omega \; \forall i, j$, them $C_i = C_i^w$.*

Proof. Following the definition of C_i^w we have:

$$C_i^w = \frac{1}{S_i(k_i-1)} \sum_{j,h} \frac{2\omega}{2} a_{jh} a_{ij} a_{ih} a_{ji} a_{hi}$$

Where $S_i = \sum_{j=1}^{|V|} a_{ij} a_{ji} \omega$. Replacing S_i, we have:

$$C_i^w = \frac{1}{\sum_{j=1}^{|V|} a_{ij} a_{ji} \omega (k_i - 1)} \sum_{j,h} \omega a_{jh} a_{ij} a_{ih} a_{ji} a_{hi}$$

$$C_i^w = \frac{1}{\sum_{j=1}^{|V|} a_{ij}a_{ji}(k_i - 1)} \sum_{j,h} a_{jh}a_{ij}a_{ih}a_{ji}a_{hi}$$

We also know that following the neighbour definition and the adjacency matrix definition: $k_i = \sum_{j=1}^{|V|} a_{ij}a_{ji} = |\Gamma_{v_i}| = |\{v_j \mid e_{ij} \in E \; and \; e_{ji} \in E\}|$ And finally:

$$C_i^w = \frac{1}{k_i(k_i - 1)} \sum_{j,h} a_{jh}a_{ij}a_{ih}a_{ji}a_{hi}$$

Which proves theorem 1

As a corollary to this theorem, if $CC_i^w = 1 \Rightarrow CC_i = 1$.

Finally, if we want to study a general graph, we should study its Global CC:

Definition 2..12 (Global CC [14, 6])**.** The clustering coefficient of a graph can be defined as:

$$C = \frac{1}{|V|} \sum_{i=0}^{|V|} C_i \tag{6}$$

Where $|V|$ is the number of vertices.

The Global Weighted Clustering Coefficient is:

$$C^w = \frac{1}{|V|} \sum_{i=0}^{|V|} C_i^w \tag{7}$$

The main difference between Local CC, Local Weighted CC and Global CC is that, the first one can be used to represent how connected is a node locally in a graph, the second one is used to calculate the density of these connections using the edge weights, and the last one provides us with global information about of the connectivity in a graph. In real complex problems only the two initial measures can be used, whereas the third one is usually estimated [57].

3. Data Mining

Data Mining is "the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" [34]. The Data Mining techniques are divided in 5 main steps:

1. **Data Extraction:** The data extraction problem consists on obtain the datasets which will be analysed.
2. **Data Preprocessing and Normalization:** The data preprocessing methods prepare the data to be analysed. There are three main steps [34]: avoid misclassification, dimensionality reduce (through projections or feature selection techniques) and range normalization.

3. **Model Generation:** This is the most important part of the data analysis. The model is created to find the patterns in the data. It is usual to use Machine Learning or other statistical techniques to generate the model [34].
4. **Model Validation:** Depending on the type of model, the validation process is different. This process gives the confidence of the model. It is usual to use validation with classifiers [34].
5. **Model Application:** The goal of the model is to be applied, for example, to predict the behaviour of new inputs.

3.1. Data Extraction

There are several public databases where the data can be taken according to the analysis goal. The most used in Data mining works is the UCI Machine Learning Repository [2] which contains several databases to test the algorithms. Also, there are some applications for Social Network analysis which allow researchers to extract information from Twitter (such as Twitter API [38]) or Facebook (Facebook API [24]), amongst others.

3.2. Data Preprocessing

Data Mining techniques need an intensive phase of data preprocessing. Initially the information must be analysed and stored in some kind of database system, cleaned and separated. This preprocessing phase is used to avoid outliers, missclassifications and missing data. Methods such as histogram and statistical correlation are used to clean the dataset and reduce the number of variables [34]. Projections are also usual in dimension reduction, however, projection methods [15] such as PCA (Principal Component Analysis) or LDA (Lineal Discriminant Analysis) do not offer a complete perspective of the problem. These methods create new variables which are estimated from principal components or lineal projections trying to separate the data and reduce its dimension. Usually, these techniques lose the original information of the features which is unrecoverable once it is projected. It produces a reduction of the human interpretation of Data Mining techniques applied and, sometimes, it is preferable to avoid them.

There are several techniques which reduce the feature sets to avoid projections. These methods apply a guided search among the different attributes looking for the most useful variables for the analysis. These methods are usually known as feature selection methods [32]. Curiel et al. [13] apply genetic algorithms to simplify prognosis of endocarditis using a codification where each individual of the population is based on a set of features. Blum and Langley [8] show some examples of relevant features selections in different datasets and applied them to different machine learning techniques. They define different degrees of relevant features such as strong or weak relevant features. They also study some methodologies such as heuristic search, filters and wrapper approaches which are automatic feature selection methods usually validated by classification techniques. Some of these techniques usually introduced over-fitting to the model and are computationally expensive. Roth and Lange [52] apply these techniques to the clustering problem.

Finally, the last step is related to normalization. It allows to compare data features with different kind of range of values. Z-Score [10] and Min-Max [26] normalization

methods are commonly used for preprocessing the data. Both normalization algorithms takes the attribute records and they find a standard range for them. Min-Max has a fixed range, [0,1] (it is sensitive to outliers), while Z-Score depends on the mean and the standard deviation (it approximates the distribution to a normal distribution, it is usually used to avoid outliers). These algorithms obtain the normalized values from data using the following equations:

- Min-max: It computes maximum and minimum values of the attributes applying:
$$x' = \frac{x - min(X)}{max(X) - min(X)}$$
- Z-Score: It computes mean and standard deviation of the values applying:
$$x' = \frac{x - mean(X)}{SD(X)}$$

Once the data is ready for the analysis, the model generation phase begins. This work is focused on unsupervised learning techniques for model generation, specially clustering techniques, that are presented in the following section.

4. Model Generation: Clustering

The most important part in the Data Mining processes is the Model Generation step where the Machine Learning algorithms are applied. In this section, several algorithms related to Graph Theory and Unsupervised Learning techniques are explained, specially one of the most important clustering algorithms related to Graph Spectrum Theory: Spectral Clustering. Finally, it presents the Community-finding problem, related to identify communities of users in the Network.

4.1. Graph Clustering

Graph theory has also proved to be an area of important contribution for research in data analysis, especially in the last years with its application to manifold reconstruction [23] using data distance and graph representation to create a structure which can be considered as an Euclidean space (which is the manifold).

Graph models are useful for diverse types of data representation. They have become especially popular over the last years, being widely applied in the Social Networks area. Graph models can be naturally used in these domains, where each node or vertex can be used to represent an agent, and each edge is used to represent their interactions. Later, algorithms, methods and Graph Theory have been used to analyse different aspects of the network, such as: structure, behaviour, stability or even community evolution inside the graph [14, 21, 41, 60].

A complete roadmap to Graph Clustering can be found in [53] where different clustering methods are described and compared using different kinds of graphs: weighted, directed, undirected. These methods are: Cutting, Spectral Analysis and Degree Connectivity (an exhaustive analysis of connectivity methods can be found in Hartuv and Shamir[27]), amongst others. This roadmap also provides an overview of computational complexity from a theoretical and experimental point of view of the

studied methods.

From previously described graph clustering techniques, a recent and really powerful ones are those based on Spectral Clustering which is introduced in the following section.

4.2. Spectral Clustering

Spectral clustering methods are based on a straightforward interpretation of weighted undirected graphs as can be seen in [1, 40, 45, 58]. The Spectral Clustering approach is based on a Similarity Graph which can be formulated in three different ways (all of them equivalent [58]) of graphs:

1. **The ϵ-neighbourhood graph**: all the components whose pairwise distance is smaller than ϵ are connected.
2. **The k-nearest neighbour graphs**: the vertex v_i is connected with vertex v_j if v_j is among the k-nearest neighbours of v_i.
3. **The fully connected graph**: all points with positive similarity are connected with each other.

The main problem is how to compute the eigenvector and the eigenvalues of the Laplacian matrix of this Similarity Graph. For example, when large datasets are analysed, the Similarity Graph of the Spectral Clustering algorithm takes too much memory, it makes difficult the eigenvalues and eigenvectors computation. Some works are focused on this problem: von Luxburg et al. [58] present the problem, Ng et al.[45] apply an approximation to a specific case, and Nadler et al.[40] apply operators to get better results. The classical algorithms can be found in [58].

The theoretical analysis of the observed good behaviour of SC is justified using the perturbation theory [58, 40], random walks and graph cut [58]. The perturbation theory also explains, through the eigengap, the behaviour of Spectral Clustering.

Some of the main problems of Spectral Clustering are related to the consistency of the two classical methods used in the analysis: normalized and un-normalized spectral clustering. A deep analysis about the theoretical effectiveness of normalized clustering over un-normalized can be found in [59].

4.2.1. The Spectral Clustering Algorithm

Spectral Clustering methods were introduced by Ng et al. in [45]. These methods apply the knowledge extracted from graph spectral theory to clustering techniques. These algorithms are divided in three main steps:

1. The algorithm constructs a graph using the data instance as nodes and applying a similarity measure to define the edges weights (see Algorithm 1 line 1). The different types of graphs are explained above. The measure which is usually used is the Radial Basis Function (RBF) Kernel (which is the most usual approach taken in the literature) defined by:

$$s(x_i, x_j) = e^{-\sigma||x_i - x_j||^2} \tag{8}$$

where σ is used to control the width of the neighbourhood.

2. It studies the graph spectrum calculating the Laplacian Matrix associated to the graph (see Algorithm 1, lines 2 and 3) . There are different definitions of the Laplacian Matrix. These definitions achieved different results when they are applied to the Spectral Clustering algorithm. They are used to categorize the Spectral Clustering techniques as follows [58]:
   - **Unnormalized Spectral Clustering** It defines the Laplacian matrix as:
     $$L = D - W \tag{9}$$
   - **Normalized Spectral Clustering** It defines the Laplacian matrix as:
     $$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \tag{10}$$
   - **Normalized Spectral Clustering (related to Random Walks)** It defines the Laplacian matrix as:
     $$L_{rw} = D^{-1} L = I - D^{-1} W \tag{11}$$

   In these formulas I is the identity matrix, D represents the diagonal matrix whose (i, i)-element is the sum of the similarity matrix ith row and W represents the Similarity Graph (see Algorithm 1 line 2). Once the Laplacian is calculated (in Algorithm 1 the Normalized Spectral Clustering algorithm is used, however, in this case, to simplify, the eigenvalues which are calculated are $1 - \lambda_i$ instead of λ_i, the eigenvectors do not change), its eigenvectors are extracted (see lines 4 and 5 of Algorithm 1).

   The three Laplacian Matrices have been deeply studied in the related literature [58, 40, 59]. They are connected to the *graph cut problem*, which looks for the best way to cut a graph keeping a high connection amongst the elements which belongs to each partition, and a low connection between the elements of different partitions.

   The graph cut problem is closely related to clustering. In the graph cut literature this problem has two classical solutions[58]: RadioCut and NCut. Von Luxburg et al. [58] describe the connection between the different approaches of SC (focused on the Laplacian Matrices), RadioCut and NCut. They also show that Unnormalized Spectral Clustering converges to RadioCut and the Normalized method converges to the NCut. On the other hand, a deep analysis about the theoretical effectiveness of Normalized clustering over Unnormalized can be found in [59].

3. The eigenvectors of the Laplacian Matrix are considered as points and a clustering algorithm, such as K-means [37], is applied over them to define the clusters (see Algorithm 1 lines 7 and 8).

4.3. Community Finding Approach

The main application of the communities approach are Social Networks. The clustering problem is more complex when is applied to find communities in networks (subgraph identifications). A community can be considered as a subset of individuals with relatively strong, direct, and intensive connections [21] between them. Some algorithms such as Edge Betweenness Centrality (EBC) [22] or Clique Percolation Method (CPM) [16] have been designed to solve this problem following a deterministic process. EBC algorithm [22] is based on finding the edges of the network which connect communities and removing them to determine a good definition of these communities. CPM [16] finds communities using k-cliques (where k is a fixed value of connections in a graph) which are defined as complete (fully connected) subgraphs of k vertices. It defines a community as the highest union of k-cliques. CPM has two

Algorithm 1 Normalized Spectral Clustering according to Ng et al. (2001)[45]

Require: A dataset of n elements $X = \{x_1, \ldots, x_n\}$ and a fix number of clusters k.
Ensure: A set of clusters $C = \{C_1, \ldots, C_k\}$ which partitionate X

1: Form the affinity matrix $W \in R^{n \times n}$ defined by $W_{ij} = e^{-||x_i - x_j||^2/2\sigma^2}$ if $i \neq j$, and $W_{ii} = 0$.
2: Define D to be the diagonal matrix whose (i, i)-element is the sum of the i-th row of W.
3: Construct the matrix $L = D^{-1/2} W D^{-1/2}$.
4: Find $v_1, \ldots, v_k$, the k largest eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues) and form the matrix $V = [v_1 v_2 \ldots v_k] \in R^{n \times k}$ by stacking the eigenvectors in columns.
5: Form the matrix Y from V by renormalizing each row of V to have unit length (i.e. $Y_{ij} = V_{ij}/(\sum_j V_{ij}^2)^{1/2}$).
6: Apply K-means (or any other algorithm) treating each row of Y as a point in R^k.
7: Assign the points x_i to cluster C_j if and only if the row i of the matrix Y was assigned to cluster j.
8: **return** C

variants: directed graphs and weighted graphs [48].

Other approximations related to the finding-community problem can be found in [51] where different statistical mechanics for community detection are used. Pons and Latapy [49] uses random walks to compute the communities. Finally, another work based on metrics used to measure the quality of the communities can be found in [44], and metrics that can be used to find the structure of a community in very large networks in [12]. Genetic algorithms have also been applied to find communities or clusters through agglomerative genetic algorithms [36] and multi-objective evolutionary algorithms [30] amongst others.

5. Complex Network Analysis

The analysis of complex networks gas become a very important field, specially in physics. It is used to analyse Social Networks which are usually represented as a Complex Network.

5.1. Types of Networks

There are four basic types of Networks which are generally considered:

- **Random Network** [19]: This network is based on random connections. Given a connection probability, the graph is usually generated asigning edges between nodes with a predefined probability which is usually small. This kind of graphs is also called as Erdős and Rényi graphs because the model to generated the graphs was introduced by these authors in 1959. One of the main properties of Random Graphs is about the Clustering Coefficient, this metric has usually small values which is not usual in real networks [4]. It gives the intuition that real

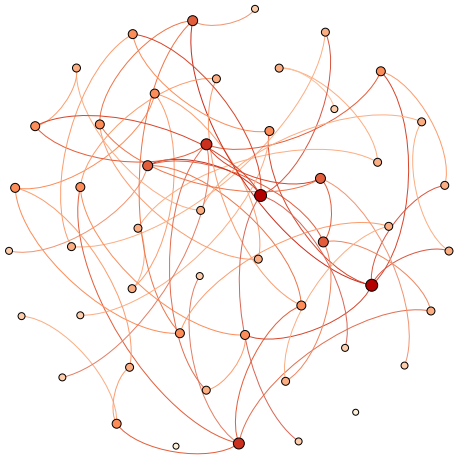


Figure 1. Example of a Random Network with 50 nodes and a connection probability of $p = 0.05$

networks have a correlation factor amongst their connections. Figure 1 shows an example of a Random Graph generated by the Erdős and Rényi using 50 nodes and a connection probability of $p = 0.05$.

- **Regular Network** [62]: This kind of networks satisfies that each vertex has the same number of neighbours. The consequence is that every node has the same degree. These networks have been studied theoretically but is difficult to find a real-world regular network. However, the Watts and Strogatz algorithm [61] takes advantage of this network structure to generate Small-World networks. An example of this kind of network can be found in Figure 2 where a regular network has been generated with 20 nodes and the degree of each node is 4.
- **Scale-Free Network** [5]: A scale-free network has a degree distribution where the probability of a node having a given degree has a scale-invariant decay as degree grows. Hence, it follows a power-law of the form

$$P(n) \sim n^{-\gamma}, \tag{12}$$

where $\gamma > 1$ is a constant and $n = 1, 2, \ldots, N$. The were introduced by Barabasi and Albert in [5] and are currently known as the BarabasiAlbert (BA) or preferential attachment model [43]. This kind of distribution is very different to that of homogeneous random networks. They are neither random nor small-world networks. The features observe in this networks are frequently observed in real-world networks. Figure 3 shows an example of this kind of networks for 45 nodes.

- **Small World Network** [61]: Small-World networks are characterized by the "the small-world effect". This term is used to describe networks whose average path length is comparable with a Regular Network without any regard to the clustering coefficient. Small-world networks were introduced by Watts and Strogatz [61]. They can be obtained via the following algorithm [14]:

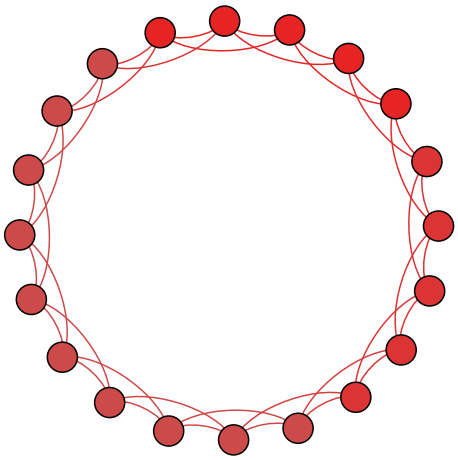


Figure 2. Example of a Regular Network with 20 nodes and where all nodes have degree 4

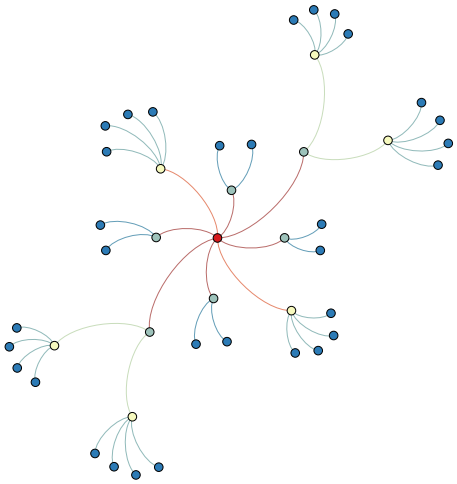


Figure 3. Example of a Scale-Free Network where $N = 45$

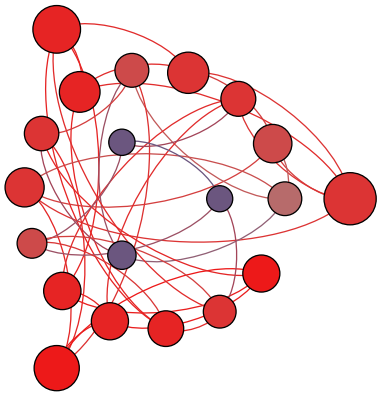


Figure 4. Example of a Small-World Network generated with the algorithm of Watts and Strogatz [61] where $p = 0.05$, $k = 4$ (neighbours), and $N = 20$ (nodes) and taking the regular graph of Figure 2 as starting point.

1. First, arrange all nodes on a ring and connect each node with its $k = 2$ nearest neighbours (see Figure 2).
2. Second, start with an arbitrary node i and rewire its connection to its nearest neighbour on, for example, the left side with probability p to any other node j in the network. Choose the next node and repeat the process.
3. Third, after all next neighbour connections have been checked repeat this procedure for the second and all higher next neighbours successively.

This algorithm guarantees that each connection occurring in the network is chosen exactly once to test for a rewiring with a fixed probability which controls the disorder of the resulting topology.

Taking a regular graph as an starting point, in which the diameter is proportional to the size of the network, it can be transformed into a "small world" in which the average number of edges between any two vertices is very small, while the clustering coefficient stays large. Figure 4 shows an example of how the Regular Network of Figure 2 can be transformed in a "Small-World" Network.

5.2. Page Rank and HITS

The analysis of a Complex or Social Network can be focused in all the information that has been mentioned above, however, there are other algorithms which deserve to be explained in this work. These algorithm are PageRank [9] and HITS [31]. They can be used to take information about the most representative nodes of the network and how they affect to it.

5.2.1. PageRank

PageRank [9] is a link analysis algorithm initially used by the Google web search engine. It assigns a numerical weigh to each element of a linked set of nodes (which

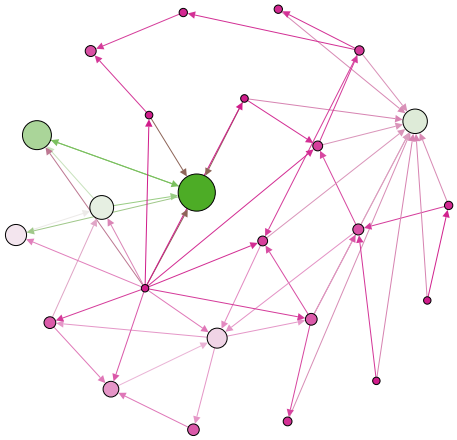


Figure 5. Example of PageRank application

in the original implementation was though as a hyperlinked set of web pages such as the World Wide Web). The purpose is to measure the importance of each node within the graph. The numerical weight assigned to each node n_i is referred to the PageRank value of n_i denoted by $PR(n_i)$.

The PageRank algorithm is an iterative algorithm which calculates recurrently the following values:

$$PR(n_i) = \frac{1-d}{N} + d \sum_{n_j \in M(n_i)} \frac{PR(n_j)}{L(n_j)} \tag{13}$$

Where $PR(n_j)$ is the PageRank value of node n_j, d is the damping factor which is used to adjusts the algorithm, N is the number of nodes, $L(n_j)$ is the number of outbounds links on node n_j and $M(n_i)$ is the set of nodes with inbound links to n_i.

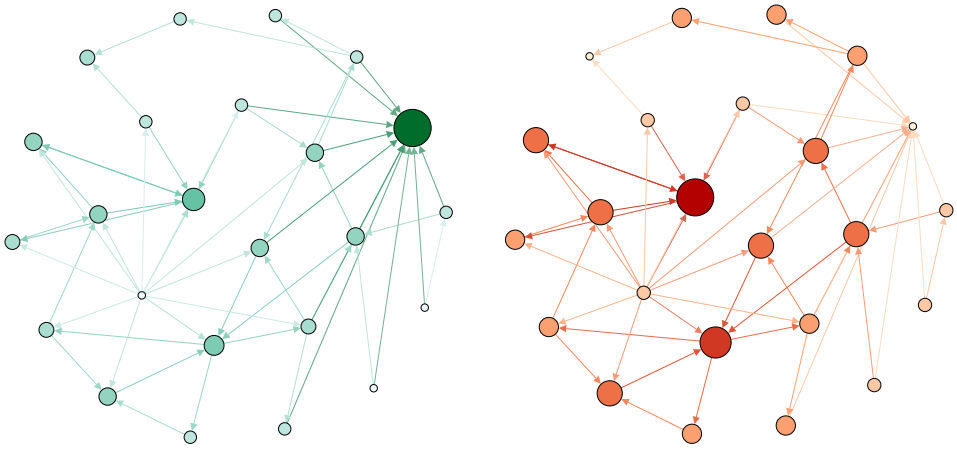
This algorithm is usually solved using an algebraic process or an iterative process. In addition, when the iterative process is used, the PageRank values are usually normalized.

Figure 5 shows an example of the PageRank application to a directed graph. The biggest nodes represents the higher PageRank values and vice-versa.

5.2.2. HITS

Hyperlink-Induced Topic Search (HITS) [31] (also known as hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It was a precursor to PageRank.

The HITS algorithm calculates two main values: hub and authority. These values



(a) Authorities (b) Hubs

Figure 6. Example of HITS algorithm application

are calculated iteratively. The authority value is calculated as follows:

$$PA(n_i) = \sum_{\{n_j | e_{ij} \in G\}} PH(n_j) \tag{14}$$

Where e_{ij} is an edge from node n_i to node n_j and G is the graph. $PH(n_j)$ is the hub value of n_j. This value is calculated as follows:

$$PH(n_i) = \sum_{\{n_j | e_{ij} \in G\}} PA(n_j) \tag{15}$$

The algorithm usually begins with $PA(n_i) = PH(n_i) = 1 \; \forall i$. Figure 6 shows an example of the HITS application to a directed graph. The biggest nodes in sub-figure a) represents the highest Authorities values while in figure b) represents the highest Hubs values. The smallest represents the lowest values respectively.

6. Applications in Social Networks Analysis

This section tries to give some practical applications for Social Data Mining. Social Analysis is one of most challenge fields since the Web 2.0 started in 1999. This type of web sites generates a more interactive source of information which promotes the interactions between the users using forums and chats originally. In 2004 Mark Zuckerberg founded Facebook, one of the most relevant Social Networks at present. Facebook allows the users to share comments and opinions. Two years later, in 2006, Jack Dorsey created Twitter. This Social Networking and Microblogging services is currently one of the most famous and challenge Social Network for text analysis.

6.1. Facebook

A social network can be analysed from different perspectives which have been described above. A good example is Facebook. Facebook is almost the most important Social Networks. It was originally created to interchange photos between users which are friends inside the Social Context, however, it currently is used for videos, messages, games, etc. The most important features of Facebook as a Social Network are:

- The 'Friendship' structure, where the users belongs to a friends community formed by people of his human context.
- The 'Like' button, which express the interest of different users in photos, videos, comments, etc, which are posted by other users or by themselves.
- The comment option which allows the users to comment everything (also comments) and generates interactions between them.

Using this structure as an starting point it is possible to analyse the Network generated. The analysis can be focused from several points of view, for example, in [20] discuss the mesoscopic features of the community structure of this network, after they unveiled the communities representing the aggregation units among which users gather and interact; they analyse the statistical features of that network communities, discovering and characterizing some specific organization patterns followed by individuals interacting in online social networks.

In [11] they focused their work on participants of online Social Networks. The data is anonymous and organized as an undirected graph. They develop a set of tools to analyse specific properties such as degree distribution, centrality measures, scaling laws and distribution of friendship.

In [3] they face a link prediction problem. Given a snapshot of a network they infer which interactions among existing members are likely to occur in the near future or which existing interactions are we missing.

Finally, [35] introduces a new public dataset based on manipulations and embellishments of Facebook. In the second half of this paper, they use a community finding algorithm to find subgroups defined by gender, race/ethnicity, and socioeconomic.

6.2. Twitter

Twitter is a Social Network where people usually publish information about personal opinions. It is divided in two kind of users behaviour: follower and following. As a follower, the user receives information of people which is followed by him, and as a following, the user information is sent to its followers. The information that the users share is called Tweets. Tweets are sentences limited by 140 characters which can contain information about personal opinions of the users, photos, links, etc. A user can also re-tweet the information of other users and share it.

The information of Twitter or other networks based on text interchange (such as forums or Facebook) can be used for analysis.

In [55] they examine the influence of geographic distance, national boundaries, language, and frequency of air travel on the formation of social ties on Twitter. They show that a substantial share of ties lies within the same metropolitan region, and

that between regional clusters, distance, national borders and language differences all predict Twitter ties.

In [29] they analyse the usage of Twitter. Also they present a taxonomy characterizing the the underlying intentions users have in making microblogging posts. By aggregating the apparent intentions of users in implicit communities extracted from the data, they show that users with similar intentions connect with each other.

In [39] they propose a method for Twitter Social Network that takes a single static snapshot of network edges and user account creation times to accurately infer when these edges were formed. This method can be exact in theory, and they demonstrate empirically for a large subset of Twitter relationships that it is accurate to within a few hours in practice.

Finally, [33] studies the topological characteristics of Twitter and its power as a new medium of information sharing. they have found a non-power-law follower distribution, a short effective diameter, and low reciprocity. In order to identify influential users on Twitter, they have ranked them by the number of followers and by PageRank and found two rankings to be similar.

7. Software Tools

There are several tools used in Data Mining and Social Data Mining analysis. Some relevant and straightforward tools are the following:

- **Gephi** [1]: Gephi [7] is a visualization and graph analysis software oriented to all kinds of networks and complex systems, dynamic and hierarchical graphs. It has several algorithms implemented for Social Network analysis such as PageRank, HITS, etc. Also it has algorithms to improve the graph visualization using different layouts and is able to calculate different metrics of the graph such as degree (power-law), betweenness, closeness, density, path length, diameter, modularity, clustering coefficient.
- **Graphviz** [2]: Graphviz [18] (Graph Visualization Software) is open source graph visualization software. It can represent diagrams, graphs and networks. It has been applied to networking, bioinformatics, software engineering, database and web design, machine learning and visual interfaces for other technical domains.
- **JUNG** [3]: JUNG [46] (the Java Universal Network/Graph Framework) is a Java library that provides some tools for graph or network modeling, analysis and visualization. It has been designed to support a variety of representations and analytic tools for complex data sets. It also provides a visualization framework with different layouts.
- **Mahout**[4]: The Apache Mahout™[47] machine learning library was designed to build scalable machine learning libraries. It has several Machine Learning tools for clustering, classfication and batch based collaborative filtering which are implemented on top of Apache Hadoop[5] using the map/reduce paradigm.

[1] https://gephi.org
[2] http://www.graphviz.org
[3] http://jung.sourceforge.net
[4] http://mahout.apache.org
[5] http://hadoop.apache.org/

- **Matlab** [6]: MATLAB®[63] is a high-level language and interactive environment for numerical computation, visualization, and programming. It has several tools for analyze data, develop algorithms, and create models and applications.
- **Octave** [7]: Octave [17] is a high-level interpreted language for numerical computations. It provides extensive graphics capabilities for data visualization and manipulation. The Octave language is similar to Matlab so that most programs are easily portable.
- **R** [8]: R [50] is a language and environment for statistical computing and graphics. It is a GNU project which provides a wide variety of statistical (such as modelling, statistical tests, time-series analysis, classification, clustering amongst others) and graphical techniques, and is highly extensible.
- **Weka** [9]: Weka [25] is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is open source software issued under the GNU General Public License.

8. Conclusions

Graph Theory have become an important tool in different Social Data Mining methods. These techniques have influenced not only in the analysis procedures such as Complex Network methods but also in the creation of new techniques such ad Spectral Clustering.

The Analysis can also be focused on both a Data Mining approach and a Complex Network approach depending of the interest of the analysis. From the Data Mining point of view, unsupervised methods such as clustering algorithm or community finding algorithms can be used to find groups of similar users within the Social Network while from the Complex Network point of view, HITS and PageRank algorithms can be used to find the most relevant users and the network analysis can be used to define the nature and robustness of the network.

Finally, these methods have several applications specially for current Social Networks such as Facebook and Twitter using different software tools developed by the research communities and companies.

References

[1] Francis Bach and Michael Jordan. Learning Spectral Clustering, With Application To Speech Separation. *Journal of Machine Learning Research*, 7:1963 – 2001, October 2006.

[2] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[3] L. Backstrom and J. Leskovec. Supervised Random Walks: Predicting and Recommending Links in Social Networks. November 2010.

[6] http://www.mathworks.co.uk/products/matlab/index.html
[7] http://www.gnu.org/software/octave
[8] http://www.r-project.org
[9] http://www.cs.waikato.ac.nz/ml/weka/

[4] A.L. Barabási. *Linked: how everything is connected to everything else and what it means for business, science, and everyday life*. Plume book. Plume, 2003.

[5] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, October 1999.

[6] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, March 2004.

[7] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.

[8] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97:245–271, December 1997.

[9] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[10] Susan Rovezzi Carroll and David J. Carroll. *Statistics Made Simple for School Leaders*. Rowman & Littlefield, 2002.

[11] Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Crawling Facebook for Social Network Analysis Purposes. pages 1+, May 2011.

[12] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111+, December 2004.

[13] Leticia Curiel, Bruno Baruque, Carlos Dueñas, Emilio Corchado, and Cristina Pérez-Tárrago. Genetic algorithms to simplify prognosis of endocarditis. In *Proceedings of the 12th international conference on Intelligent data engineering and automated learning*, IDEAL'11, pages 454–462, Berlin, Heidelberg, 2011. Springer-Verlag.

[14] M. Dehmer, editor. *Structural Analysis of Complex Networks*. Birkhäuser Publishing, 2010. in press.

[15] K. Delac, M. Grgic, and S. Grgic. Independent comparative study of PCA, ICA, and LDA on the FERET data set. *International Journal of Imaging Systems and Technology*, 15(5):252–260, 2005.

[16] Imre Derényi, Gergely Palla, and Tamás Vicsek. Clique Percolation in Random Networks. *Physical Review Letters*, 94(16):160202–1 – 160202–4, Apr 2005.

[17] John W. Eaton. *GNU Octave Manual*. Network Theory Limited, 2002.

[18] John Ellson, Emden R. Gansner, Eleftherios Koutsofios, Stephen C. North, and Gordon Woodhull. Graphviz - Open Source Graph Drawing Tools. *Graph Drawing*, pages 483–484, 2001.

[19] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[20] Emilio Ferrara. A Large-Scale Community Structure Analysis In Facebook, March 2012.

[21] Santo Fortunato, Vito Latora, and Massimo Marchiori. Method to find community structures based on information centrality. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 70(5):056104, 2004.

[22] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, June 2002.

[23] Alexander N. Gorban and Andrei Zinovyev. Principal manifolds and graphs in practice: From molecular biology to dynamical systems. *International Journal of Neural Systems*, 20(3):219 – 232, 2010.

[24] W. Graham. *Facebook API Developers Guide*. Apresspod Series. Apress, 2008.

[25] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[26] Jiawei Han and Micheline Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.

[27] Erez Hartuv and Ron Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4–6):175–181, 2000.

[28] D. Helbing and S. Balietti. From social data mining to forecasting socio-economic crises. *The European Physical Journal - Special Topics*, 195(1):3–68, May 2011.

[29] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why We Twitter: An Analysis of a Microblogging Community. In Haizheng Zhang, Myra Spiliopoulou, Bamshad Mobasher, Giles, Andrew McCallum, Olfa Nasraoui, Jaideep Srivastava, and John Yen, editors, *Advances in Web Mining and Web Usage Analysis*, volume 5439 of *Lecture Notes in Computer Science*, chapter 7, pages 118–138. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[30] Keehyung Kim, RI (Bob) McKay, and Byung-Ro Moon. Multiobjective evolutionary algorithms for dynamic social network clustering. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, GECCO '10, pages 1179–1186, New York, NY, USA, 2010. ACM.

[31] Jon M. Kleinberg. Hubs, authorities, and communities. *ACM Comput. Surv.*, 31(4es), December 1999.

[32] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97:273–324, December 1997.

[33] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[34] Daniel T. Larose. *Discovering Knowledge in Data*. John Wiley and Sons, 2005.

[35] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4):330–342, October 2008.

[36] Marek Lipczak and Evangelos Milios. Agglomerative genetic algorithm for clustering in social networks. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, GECCO '09, pages 1243–1250, New York, NY, USA, 2009. ACM.

[37] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[38] Kevin Makice. *Twitter API: Up and Running: Learn How to Build Applications with the Twitter API*. O'Reilly Media, Inc., 1 edition, April 2009.

[39] Brendan Meeder, Brian Karrer, Amin Sayedi, R. Ravi, Christian Borgs, and Jennifer Chayes. We know who you followed last summer: inferring social link creation times in twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 517–526, New York, NY, USA, 2011. ACM.

[40] Boaz Nadler, Stéphane Lafon, Ronald Coifman, and Ioannis G. Kevrekidis. Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators. pages 955 – 962, 2005.

[41] Mariá Cristina Vasconcelos Nascimento and André C. P. L. F. Carvalho. A graph clustering algorithm based on a clustering coefficient for weighted graphs. *J. Braz. Comp. Soc.*, 17(1):19–29, 2011.

[42] Tamás Nepusz. *Data mining in complex networks: Missing link prediction and fuzzy communities*. PhD thesis, Budapest University of Technology and Economics, 2008.

[43] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003.

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

[45] A. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001.

[46] J. O'Madadhain, D. Fisher, S. White, and Y. Boey. The JUNG (Java Universal Network/Graph) Framework. Technical report, UCI-ICS, October 2003.

[47] Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman. *Mahout in Action*. Manning Publications, 1 edition, January 2011.

[48] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.

[49] P. Pons and M. Latapy. Computing communities in large networks using random walks (long version). *ArXiv Physics e-prints*, December 2005.

[50] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[51] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, Jul 2006.

[52] Volker Roth and Tilman Lange. Feature selection in clustering problems. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[53] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[54] Tom Smith. The social media revolution. *International Journal of Market Research*, 51(4):559–561, July 2009.

[55] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of Twitter networks. *Social Networks*, 34(1):73–81, January 2012.

[56] Lindsay A. Thompson, Kara Dawson, Richard Ferdig, Erik W. Black, J. Boyer, Jade Coutts, and Nicole Paradise P. Black. The intersection of online social networking with medical professionalism. *Journal of general internal medicine*, 23(7):954–957, July 2008.

[57] A Tsonis, K Swanson, and G Wang. Estimating the clustering coefficient in scale-free networks on lattices with local spatial correlation structure. *Physica A: Statistical Mechanics and its Applications*, 387(21):5287–5294, 2008.

[58] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[59] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, April 2008.

[60] D. J. Watts. *Small worlds : the dynamics of networks between order and randomness*. 1999.

[61] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–10, 1998.

[62] Waigai Zhen. *Graph Theory and its Engineering Applications.* Advanced series in electrical and computing engineering. World Scientific Publishing Company, Incorporated, 1997.

[63] Dongli Zhou, Wesley K. Thompson, and Greg Siegle. MATLAB toolbox for functional connectivity. *NeuroImage*, 47(4):1590–1607, October 2009.