Restoring Information Needed for Social Internetworking Analysis from Anonymized Data

Francesco Buccafurri, Daniele Caridi, Gianluca Lax, Antonino Nocera and Domenico Ursino
Dept. of Information, Infrastructure and Sustainable Energy Engineering
University “Mediterranea” of Reggio Calabria
Reggio Calabria, Italy
Email: {bucca,daniele.caridi,lax,a.nocera,ursino}@unirc.it

***Abstract*—The interaction among distinct social networks is the basis of a new emergent internetworking scenario (called *Social Internetworking Scenario* or, simply, SIS), enabling a lot of strategic applications whose main strength will be just the integration of possibly different communities yet preserving their diversity and autonomy. As a consequence, studying this new scenario from a Social-Network-Analysis perspective is certainly an important and topical issue, also for the possibility of discovering a lot of relevant knowledge about multiple aspects of people life. However, not always the analyst is able to deal with the hard problem of collecting data through the execution of a crawler. In this case, she could exploit graph-based social data, collected by another party, and usually anonymized for privacy reasons. Unfortunately, even the most frequent and trivial anonymization (i.e., the elimination of URLs associated to nodes), handicaps a lot of SIS-oriented investigations, due to the lack of some relevant explicit information. In this paper, we deal with this problem, by proposing and by experimentally validating a clustering-based technique able to restore part of the missing explicit information, thus allowing the profitable analysis of anonymized multi-social-network data.**

***Keywords*-Social Network; Social Network Analysis; Social Internetworking System; Anonymized Data; Clustering**

I. Introduction

In the last years, (on line) social networks have become one of the main actors not only of the Cyberspace but also of real life. Indeed, an always increasing number of persons joins one or more social networks, and large areas of economy, politics, communication and, more in general, of all the aspects of real life, make a large use of this innovative tool. The extraordinary development of this phenomenon has led to the presence of several different social networks, whose number is expected to increase, each incorporating peculiar features making it more or less attractive for the market. Thus, a user joins Facebook to talk to her friends, YouTube to share her videos, Flickr to store her photos, LinkedIn to look for a job, and so on. The resulting scenario is not the one of single, isolated, independent social networks, but a universe composed of a constellation of several social networks, each forming a community with specific connotations, but strongly interconnected with each other.

It is a matter of fact that, despite the inherent underlying heterogeneity, the interaction among distinct social networks is the basis of a new emergent internetworking scenario enabling a lot of strategic applications, whose main strength will be just the integration of possibly different communities yet preserving their diversity and autonomy. Clearly, social mining and analysis approaches should not miss this huge multi-network source of information, which also reflects multiple aspects of people personal life [19], thus enabling a lot of powerful discovering activities. As a matter of fact, this scenario represents the natural substrate for the development of multi-social-network applications, following a process that is already started (for instance, think of Google Open Social [4], Power.com [5], Gathera [3] and Friendfeed [2]), and will lead to increasingly powerful and innovative systems, thus increasing the importance of the so called field of *social internetworking systems* [28], [14], and determining the raising attraction of both research and industry.

Classical social networks have been studied since several years, initially by sociologists and, then, with the advent of on line social networks, by computer science researchers [20]. Studying *Social Internetworking Scenarios* (SIS’s, for short) [9], [11] means analyzing specific aspects (mainly concerning the interconnection among different social networks [10]) and adopting investigation methodologies which cannot be trivially derived from single-social-network analyses. However, the starting point is always the same.

In order to carry out meaningful analyses, we must have the availability of sufficiently large social dataset, extracted from real-life social networks, including as much information as possible. However, not always the analyst is able to deal with the hard problem of collecting data through the execution of a crawler. In fact, this is a very time-consuming activity also requiring highly performing hardware. In this (frequent) case, she could exploit data made available on the Internet by researchers who collected them in the past. For instance, Mislove at al. [26] make available a set of 11 millions users and 328 millions links among them referred to four different social networks. As it typically happens also in the context of data mining, the party that conducts the analysis on data is different from the party that collects them.

Hence, the party that exports data has also to deal with privacy issues, by purifying data from information which can be related to individuals. In the context of social network analysis, this means that collected graph-data, which represent friendships among different users, do not allow a node of the graph to be related with a real user of the social network [27]. Thus, URLs associated to nodes cannot be given.

Observe that an anonymization based on an alphabetic substitution encryption done on URLs is not secure, because it is well known that alphabetic substitution is vulnerable w.r.t. frequency-analysis attacks. Therefore, in privacy-aware contexts, anonymized social data simply consist in a graph, with no information about the URLs associated to nodes. However, in a SIS context, working with such anonymized data results in the lost of relevant explicit information. As a matter of fact, the evidence of the existence of different social networks and their interconnections is lost. Clearly, this problem does not occur whenever a single social network is investigated.

In this paper, we deal with the issue of restoring part of the missing explicit information, crucial for SIS-oriented analysis, by partitioning the whole graph in subgraphs, each corresponding as much as possible to an original social network, and, as difference, by discovering the interconnections among social networks. Once this task has been done, the analyst will be able to conduct most of the SIS-oriented investigations in a direct way, starting from the basic information concerning the membership of two or more users either to the same or to distinct social networks.

In order to carry our investigation, it appears reasonable to apply clustering techniques. Clustering represents one of the most important sectors of Data Mining. It aims at grouping a set of objects into homogeneous groups called clusters. In the Data Mining context, clustering has been exploited in a very large number of application contexts, and often extremely important results have been obtained. Despite the seemingly high execution time, clustering has been extensively exploited to analyze single social networks [34], [15], [12], [30], [22], [18], [31], [35]. Indeed, there exists a great number of clustering techniques specifically conceived to operate on large datasets. However, to the best of our knowledge, it was never adopted to investigate SIS's. Nevertheless, the intrinsic nature of a SIS makes it well suited to be investigated by means of clustering. As a matter of fact, a SIS can be seen as a set of social networks which cooperate each other. Each social network can be seen as a cluster. Therefore, the analysis of the features of a cluster can contribute to define the features of the corresponding social network.

In this paper, we verify the effectiveness of this idea. For this purpose, first we applied a crawling technique (namely, BFS [23]) on the SIS to extract a sample of data. The considered SIS consists of 5 social networks, namely Twitter, YouTube, Flickr, MySpace, LiveJournal. We chose these social networks because they are compliant with FOAF [8] and XFN [7] standards. These allow a crawler to access some not private information stored in the corresponding social networks. Then, we anonymized these data and we applied different clustering techniques on them. Finally, we compared obtained clusters with the original social networks to test the accuracy of clustering techniques in finding clusters coincident with the original social networks. In this activity, we did not exploit anonymized data available on the Web since they did not report the social network associated to each anonymized node and, hence, they were not able to allow us to measure the accuracy of the clustering techniques.

This paper is organized as follows: in Section II, we examine related literature. In Section III, we illustrate in detail the experimental analysis aimed at testing the capability of some clustering techniques to obtain clusters coincident with the original social networks of a SIS when data are anonymized. Finally, in Section IV, we draw our conclusions.

II. Related Work

In the context of Social Network Analysis [12], many of the approaches follow techniques based on clustering, analogously to what we propose in this paper. As a matter of fact, despite the seemingly high execution time, clustering has been extensively exploited to analyze single social networks. In this section, we survey the main contributions that clustering techniques have given to the investigation of social networks.

In [34], social network users and their relationships are modeled by means of weighted graphs, and a suitable measure of the density of a sub-graph, which is an index of user correlation, is defined. The proposed density measure is used to decide whether adding a node to a cluster. The clustering task of this approach aims at detecting community structures in a social network.

The authors of [15] use a genetic algorithm as a heuristic search technique to cluster social networks. This algorithm allows the social network graph to be represented in a succinct way, thus increasing the efficiency of the clustering algorithm. In order to identify clusters, it adopts a distance measure based on random walks. The results of the experimental evaluation show that the adoption of random-walk-based distances, instead of the Euclidean ones, returns more accurate clusters. [12] proposes a methodology to discover possible aggregations of nodes covering specific positions in a graph (e.g., central nodes), as well as very relevant clusters.

In [30], the authors carry out a study of temporal relation co-clustering on directional social network and author-topic evolution. Interestingly enough, the corresponding approach includes the time dimension typically ignored by the other related approaches. The analysis performed by these authors obtains meaningful results about evolution patterns, the evolution of publication topics and the evolution of e-mail communication patterns over time.

The authors of [22] developed a model and a Bayesian technique to infer four features in a social network. These features are: *(i) transitivity* - if A relates to B and B to C, then A is more likely to relate to C; *(ii) homophily* - nodes with similar characteristics are more likely to be related; *(iii) clustering into groups* - ties are more dense within groups than between them; - *(iv) degree heterogeneity* - the tendency of some actors to send and/or receive links more than others.

The authors of [18] compare three approaches devoted to engineer a hierarchical ontology on the basis of user interests in a social network. The first approach uses Wikipedia to find interest definitions, the latent semantic analysis technique to measure the similarity between interests based on their definitions, and an agglomerative clustering algorithm to group similar interests into higher level concepts. The second one uses the Wikipedia Category Graph to extract relationships between interests. The third one uses Directory Mozilla to extract relationships between interests. The results of the authors' investigations show that, although the third approach is the simplest one, it is the most effective in building a hierarchy of user interests.

In [31], the authors investigate the use of data mining techniques to identify intra- and inter-organization clusters of people with similar profiles that could have relationships among them. The proposed approach exploits a clustering method, along with a link mining-based technique that uses the minimum spanning tree to construct group hierarchies. The authors performed their analyses on a scientific social network in Brazil; in this network two scientists are connected if they have co-authored a paper. As a result of their investigation, the authors show that it is possible to derive relationships between educational intuitions by analyzing the relationships among the scientists working in them.

The authors of [35] propose the use of automatic text analysis for clustering a social network. They applied their technique to the values contained in the Enron e-mail dataset [1] and obtained the following results: *(i)* individuals communicate more frequently with individuals sharing similar value patterns than with individuals having different value patterns; *(ii)* people who communicate more frequently with each other do not necessarily all fit into a particular value type.

All the above techniques discovering social communities, despite their relationship with our approach, are not comparable with it, because they need a number of information we do not assume available in anonymized data.

III. Deriving social networks from anonymized data

In this section, we describe in detail our research efforts devoted to verify the effectiveness of applying clustering techniques to derive social networks from anonymized data. This section is organized as follows: first, we describe the testbed. Then, we discuss the issues concerning the construction of the Dissimilarity Matrix, which plays a key role in clustering techniques. Finally, we compare several clustering techniques against their capability of supporting our goal.

A. Experimental settings

As pointed out in the introduction, in order to perform our analyses, we had to extract some samples from a SIS. This last one contained social networks compliant with XFN and FOAF, which are two standards encoding human relationships in social networks. XFN [7] simply uses an attribute, called `rel`, to specify the kind of relationship between two users. Some possible values of `rel` are `me`, `friend`, `contact`, `co-worker`, `parent`, and so on. In particular, `rel` set to `me` denotes the presence of a `me` edge, which is an edge exploited to link two accounts of the same user in two different social networks. A (presumably) more complex alternative to XFN is FOAF (Friend-Of-A-Friend) [8]. A FOAF profile is essentially an XML file describing a person, her links with other people and the links to the objects created by her. It is worth pointing out that the technicalities concerning these two standards are not to be handled manually by the user. As a matter of fact, each social network has suitable mechanisms to automatically manage them in a way transparent to the user, who has just to specify her relationships in a user-friendly fashion.

The social networks of the SIS into consideration were five, namely Twitter, YouTube, Flickr, MySpace and LiveJournal. We choose these five social networks because they have been largely analyzed in the past in Social Network Analysis papers devoted to study a single social network or to compare different social networks [21], [25], [13], [33].

The crawling technique we adopted was Breadth First Search (BFS, for short). It is the most popular and widely used strategy for performing topology measurements at several levels, including sampling large networks, and is preferred with respect to other graph traversal techniques, like Depth-First Search, Forest Fire and Snowball Sampling [23]. The crawled samples presented both connections between the accounts of different users of the same social network and connections between the accounts of the same users in different social networks. They can be represented by a direct graph whose nodes correspond to user accounts and whose edges correspond to the connections between these accounts.

Once samples were obtained, they were anonymized by replacing each URL with a numeric identifier unique for all the SIS. After this, several clustering techniques were applied on anonymized samples and the clusters obtained by these techniques were compared with the original social networks. As for the adopted clustering techniques, three of them (i.e., SimpleKMeans, EM and Hierarchical) are classical and well known in literature [17]. The fourth one, called Sequential Information Bottleneck (SIB, for short) integrates an agglomerative procedure in a sequential clustering algorithm [29]. In our experiments, we set the number of clusters to be found to 5, i.e., to the number of the social networks of the SIS. Observe that this fact does not represent a real limitation since, even when the data of a SIS are anonymized, it is always specified which are the social networks composing the SIS even if the membership of a node to a certain social network has been lost during the anonymization process. We adopted the implementations of the considered clustering techniques provided by WEKA [6].

For our experiments, we exploited a server equipped with a 2 Quad-Core E5440 processor and 16 GB of RAM with the CentOS 6.0 Server operating system. We performed the crawling tasks from February 5, 2012 to March 20, 2012. After this task, we performed the other activities described above.

We considered 10 samples for this analysis; each sample referred to 5000 visited nodes.

As a first quantitative index to evaluate the effectiveness of applying clustering techniques to derive social networks from anonymized data, we adopted the Jaccard coefficient, which is a statistical index used to measure the similarity and the diversity of two sets. Given two sets A and B, the Jaccard coefficient is defined as $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ [24].

As a second quantitative index, we adopted an index called Average Cluster Partitioning (ACP) which operates at a higher abstraction level w.r.t. the Jaccard coefficient. Specifically, ACP is computed by averaging the maximum Jaccard Coefficient of each obtained cluster.

Both indexes range in the real interval $[0, 1]$. The higher their value, the better the performance of the corresponding clustering technique.

B. Dissimilarity Matrix construction

In order to apply the clustering techniques, we had to construct the Dissimilarity Matrix. In fact, it is the structure which most of the clustering techniques operate on. This matrix is $n \times n$, where n is the number of the graph nodes. Its generic element (i, j) represents the dissimilarity degree between the nodes i and j. As for the dissimilarity measure to exploit for the construction of the Dissimilarity Matrix, we chose the minimum distance between the corresponding nodes of the graph. This distance was computed by fixing the weight of each edge equal to 1 and by exploiting the Dijkstra Algorithm for the computation of the minimum paths in the graph. If it did not exist a path between two nodes, the corresponding distance was set equal to ∞.

As for this matrix, two important decisions must be made, namely: *(i)* it must be symmetric or asymmetric? *(ii)* how to normalize it in such a way that its elements range in the real interval $[0, 1]$ (i.e., in the form taken in input by clustering techniques)?

In order to answer the first question, we made the following reasoning: four of the five social networks belonging to our SIS have asymmetrical links; this makes it natural to adopt an asymmetric Dissimilarity Matrix.

In order to answer the second question, we considered three different normalization strategies which weight the absence of a path between two nodes in a different way.

In the first one, we computed the maximum among the lengths of the minimum paths between the pairs of the nodes of the graph and we divided each matrix element by this maximum incremented of 1 (this length ranged from 15 to 20 in the various samples). When the distance between two nodes was ∞, the corresponding matrix element was set to 1. This way, the normalized coefficients of the Dissimilarity Matrix for a pair of nodes not linked by a path is comparable with the normalized coefficients of the pair of nodes linked by the path with the maximum length. This choice aimed at not excessively penalizing the nodes not linked by any path. It was motivated by the idea that, actually, two nodes whose minimum path has a high length were not in a situation very different from the one of two nodes not linked by any path. This idea is also substantiated by sociological theories (see, for instance, the six-degree of separation and the small-world theories [32]) As for this normalization strategy, the ACP values obtained by applying SimpleKMeans, EM, Hierarchical and SIB were 0.762, 0.812, 0.622 and 0.760, respectively.

In the second normalization strategy, we divided the coefficients corresponding to nodes linked by paths by 100 (i.e., by a value much higher than the length of the maximum path, which ranged from 15 to 20), whereas we set to 1 the coefficients corresponding to nodes not linked by a path. In this way, we established a significant difference between the coefficients associated with the nodes linked by the maximum path and those ones corresponding to nodes not linked by a path. At the end of this experiment, the ACP values obtained by applying SimpleKMeans, EM, Hierarchical and SIB were 0.652, 0.812, 0.582 and 0.654, respectively. From the analysis of these values we obtained that this new normalization strategy negatively influenced the results of SimpleKMeans, Hierarchical and SIB, whereas the results produced by EM were identical to the ones obtained with the previous normalization strategy. As a consequence, on the whole, this last normalization strategy appears worse than the previous one.

In the third normalization strategy, we repeated this experiment by dividing the coefficients corresponding to nodes linked by a path by 250, instead of by 100, in such a way as to obtain a more significant difference between the coefficients associated with nodes linked by the maximum path and the coefficients corresponding to nodes not linked by a path. Obtained ACP values for SimpleKMeans, EM, Hierarchical and SIB were 0.621, 0.812, 0.567 and 0.622, respectively. At the end of this experiment we observed that SimpleKMeans, Hierarchical and SIB produced worse results with respect to the corresponding ones returned by the first two strategies. EM, instead, obtained the same results as the ones of the previous cases; this implies that it is not influenced by the normalization strategy.

At the end of this experiment, we concluded that the best normalization strategy was the first one.

In order to verify if our conjecture about the decision *(i)* was correct, we applied the symmetric Dijkstra algorithm, instead of the classical one, in order to obtain a *symmetric* Dissimilarity Matrix, as it generally happens for the input of clustering techniques. In particular, for each pair of nodes n_a and n_b, we computed the corresponding value of the Dissimilarity Matrix by taking the minimum between the distance from n_a to n_b and the one from n_b to n_a. Clearly, in order to normalize the Dissimilarity Matrix, we chose the first strategy. Once the Dissimilarity Matrix was constructed, we applied the four clustering techniques. Finally, we computed the corresponding ACP values and we obtained 0.382 for SimpleKMeans, 0.700 for EM, 0.486 for Hierarchical and 0.498 for SIB. From the analysis of these values it is possible to conclude that the adoption of a symmetric Dissimilarity Matrix produces worse results than the ones returned by the adoption of an asymmetrical Dissimilarity Matrix. Interestingly enough, choosing an asymmetric Dissimilarity Matrix is not the standard choice to adopt in clustering. As a matter of fact, Dissimilarity Matrixes provided in input to clustering techniques are generally symmetric. Our reasoning about the characteristics of the involved social networks allowed us to make the right decision.

C. Clustering technique comparison

The first clustering technique we applied was SimpleKMeans. For each sample, we computed the Jaccard coefficient between each cluster generated by SimpleKMeans and each social network of the SIS. Obtained results, averaged across all samples, are reported in Table I. From the analysis of this table, we can observe a correlation between clusters and social networks. In fact, clusters are capable of identifying the involved social networks. In particular, Cluster 3 perfectly corresponds to MySpace since the associated Jaccard coefficient is 1. In other three clusters, namely Clusters 1, 2 and 4, only nodes belonging to a single social network are contained. Only Cluster 0 contains all the nodes of YouTube but also nodes of Twitter, LiveJournal and Flickr. At the end of this analysis, we may conclude that SimpleKMeans finds quite a good correspondence between social networks and clusters.

TABLE I. Jaccard coefficients regarding the clusters obtained by SimpleKMeans.

	Flickr	YouTube	MySpace	LiveJournal	Twitter
Cluster 0	0.01	0.55	0.00	0.21	0.10
Cluster 1	0.00	0.00	0.00	0.00	0.76
Cluster 2	0.00	0.00	0.00	0.54	0.00
Cluster 3	0.00	0.00	1.00	0.00	0.00
Cluster 4	0.96	0.00	0.00	0.00	0.00

TABLE II. Jaccard coefficients regarding the clusters obtained by EM.

	Flickr	YouTube	MySpace	LiveJournal	Twitter
Cluster 0	0.96	0.00	0.00	0.00	0.00
Cluster 1	0.01	0.80	0.00	0.01	0.05
Cluster 2	0.00	0.00	0.00	0.00	0.76
Cluster 3	0.00	0.00	1.00	0.00	0.00
Cluster 4	0.00	0.00	0.00	0.54	0.00

TABLE III. Jaccard coefficients regarding the clusters obtained by Hierarchical.

	Flickr	YouTube	MySpace	LiveJournal	Twitter
Cluster 0	0.01	0.31	0.00	0.16	0.45
Cluster 1	0.96	0.00	0.00	0.00	0.00
Cluster 2	0.00	0.00	0.00	0.54	0.00
Cluster 3	0.00	0.00	1.00	0.00	0.00
Cluster 4	0.00	0.16	0.00	0.00	0.00

TABLE IV. Jaccard coefficients regarding the clusters obtained by SIB.

	Flickr	YouTube	MySpace	LiveJournal	Twitter
Cluster 0	0.01	0.54	0.00	0.21	0.10
Cluster 1	0.00	0.00	0.00	0.00	0.76
Cluster 2	0.00	0.00	0.00	0.54	0.00
Cluster 3	0.00	0.00	1.00	0.00	0.00
Cluster 4	0.96	0.00	0.00	0.00	0.00

We have then applied EM to the same samples and we have computed the Jaccard coefficient between each returned cluster and each social network. Obtained results are reported in Table II. From the analysis of this table, it is possible to see that the results returned by EM are better than the ones returned by SimpleKMeans. Indeed, there is an optimal correspondence between clusters and social networks.

When we applied Hierarchical to the same samples we obtained worse results than the ones returned by SimpleKMeans and EM. The corresponding Jaccard coefficients are shown in Table III. This algorithm behaves as the previous ones as far as MySpace, Flickr and LiveJournal are concerned, whereas it shows a worse behavior for the other two social networks. In particular, in Cluster 4, only nodes of YouTube occur, but the Jaccard coefficient between YouTube and Cluster 4 is significantly lower than the one between YouTube and Cluster 0. Hierarchical tends to put the nodes of YouTube and Twitter in the same cluster. We may conclude that the clusters generated by this algorithm reflect the structure of the SIS in a less precise way than the clusters returned by SimpleKMeans and EM. It seems that Hierarchical is incapable of distinguishing clusters in presence of a certain number of `me` edges (see below). This can be explained by considering that Hierarchical is well suited when it is necessary to construct cluster hierarchies by proceeding in an agglomerative fashion. In our scenario, we would have a one-level hierarchy. Furthermore, the presence of `me` edges would make it very difficult to proceed in an agglomerative fashion because the corresponding aggregation process would be quite irregular with frequent hops from a social network to another.

The last clustering technique we applied is SIB. In Table IV, we report the Jaccard coefficients between the social networks of the SIS and the clusters returned by this algorithm. We may observe that obtained clusters are equivalent to the ones returned by SimpleKMeans.

At the end of this analysis, it emerged that the best clustering technique is EM. This is also confirmed by all the values of ACP obtained in the experiments shown in Section III-B. However, the analysis above was useful because, with respect to ACP values, Tables I - IV provide more detailed results.

IV. Conclusion and Future Work

In this paper, we have studied the effectiveness of deriving social networks from anonymized data. We have explained that this problem is very important when data exploited for the analysis are taken from publicly available repositories. In this case, a lot of data are available but they have been anonymized in order to protect user privacy. The motivation of the work is that the derivation of social networks from anonymized data is needed for Social Internetworking Analysis.

As a future work, we plan to extend our research efforts in several directions. First of all, we observe that this paper is the first attempt of reconstructing information from anonymized SIS data. In order to facilitate our tasks we considered not very large samples, even though in line with those exploited by clustering techniques applied to Social Network scenarios in the past literature. We plan to analyze much larger samples in the future also considering clustering techniques specifically conceived for large datasets [16] as well as incremental clustering techniques [17]. Finally, we plan to define approaches for SIS analysis based on other Data Mining tasks (association rule extraction, classification and, above all, outlier analysis).

Acknowledgment

This work has been partially supported by the TENACE PRIN Project (n. 20103P34XC) funded by the Italian Ministry of Education, University and Research.

References

[1] Enron email dataset, http://www.cs.cmu.edu/∼enron [retrieved: May, 2013].

[2] FriendFeed, http://friendfeed.com [retrieved: May, 2013].

[3] Gathera, http://www.gathera.com [retrieved: May, 2013].

[4] Google Open Social, http://code.google.com/intl/it-IT/apis/opensocial [retrieved: May, 2013].

[5] Power.com, http://techcrunch.com/2011/04/21/power-com-shuts-down-domain-name-up-for-sale [retrieved: May, 2013].

[6] WEKA - Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka [retrieved: May, 2013].

[7] XFN - XHTML Friends Network, http://gmpg.org/xfn [retrieved: May, 2013].

[8] D. Brickley and L. Miller, "The Friend of a Friend (FOAF) project," http://www.foaf-project.org [retrieved: May, 2013].

[9] F. Buccafurri, V.D. Foti, G. Lax, A. Nocera, and D. Ursino, "Bridge Analysis in a Social Internetworking Scenario," Information Sciences, Elsevier, vol. 224, March 2013, pp. 1–18.

[10] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino, "Discovering Links among Social Networks," Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012), Springer, September 2012, pp. 467–482.

[11] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino, "Crawling Social Internetworking Systems," Proc. International Conference on Advances in Social Analysis and Mining (ASONAM 2012), IEEE/ACM, August 2012, pp. 505–509.

[12] P. Carrington, J. Scott, and S. Wasserman, Models and Methods in Social Network Analysis. Cambridge University Press, 2005.

[13] X. Cheng, C. Dale, and J. Liu, "Statistics and Social Network of Youtube Videos," Proc. International Workshop on Quality of Service (IWQoS 2008), IEEE press, June 2008, pp. 229–238.

[14] P. De Meo, A. Nocera, G. Terracina, and D. Ursino, "Recommendation of similar users, resources and social networks in a Social Internetworking Scenario," Information Sciences, Elsevier, vol. 181(7), April 2011, pp. 1285–1305.

[15] A. Firat, S. Chatterjee, and M. Yilmaz, "Genetic clustering of social networks using random walks," Computational Statistics and Data Analysis, Dec. 2007, vol. 51, pp. 6285–6294.

[16] V. Ganti, R. Ramakrishnan, J. Gehrke, A.L. Powell, and J.C. French, "Clustering Large Datasets in Arbitrary Metric Spaces," Proc. IEEE International Conference on Data Engineering (ICDE'99), IEEE Press, March 1999, pp. 502–511.

[17] J. Han and M. Kamber, Data Mining: Concepts and Techniques - Second Edition. Morgan Kaufmann Publishers, 2006.

[18] M. Haridas and D. Caragea, "Exploring Wikipedia and DMoz as Knowledge Bases for Engineering a User Interests Hierarchy for Social Network Applications," Proc. International Workshop On the Move to Meaningful Internet Systems (OTM 2009), Springer Press, Nov. 2009, pp. 1238–1245.

[19] OFCOM The independent regulator and competition authority for the UK communications industries, "Social Networking: A quantitative and qualitative research report into attitudes, behaviours and use". http://stakeholders.ofcom.org.uk/binaries/research/media-literacy/report1.pdf [retrieved: May, 2013].

[20] J. Kleinberg, "The convergence of social and technological networks," Communications of the ACM, vol. 51, Nov. 2008, pp. 66–72.

[21] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about Twitter," Proc. First Workshop on Online Social Networks (WOSN 08), ACM Press, Aug. 2008, pp. 19–24.

[22] P.N. Krivitsky, M.S. Handcock, A.E. Raftery, and P.D. Hoff, "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," Social Networks, vol. 31(3), 2009, pp. 204–213.

[23] M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of BFS (Breadth First Search)," Proc. IEEE International Teletraffic Congress (ITC 22), IEEE Press, Sept. 2010, pp. 1–8.

[24] C.D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge University Press, 2008.

[25] A. Mislove, H.S. Koppula, K.P. Gummadi, F. Druschel, and B. Bhattacharjee, "Growth of the Flickr Social Network," Proc. First Workshop on Online Social Networks (WOSN 08), ACM Press, Aug. 2008, pp. 25–30.

[26] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," Proc. SIGCOMM International Conference on Internet Measurement (IMC'07), ACM Press, October 2007, pp. 29–42.

[27] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," Proc. IEEE International Symposium on Security and Privacy (S&P 2009), IEEE Press, May 2009, pp. 173–187.

[28] Y. Okada, K. Masui, and Y. Kadobayashi, "Proposal of Social Internetworking," Proc. International Human.Society@Internet Conference (HSI 2005), Springer Press, July 2005, pp. 114–124.

[29] J. Peltonen, J. Sinkkonen, and S. Kaski, "Sequential information bottleneck for finite data," Proc. International Conference on Machine learning (ICML'04), ACM Press, July 2004, pp. 82–88.

[30] W. Peng and T. Li, "Temporal relation co-clustering on directional social network and author-topic evolution," Knowledge and Information Systems, Springer, 2011, vol. 26(3), pp. 467–486.

[31] V. Ströele et al., "Mining and Analyzing Organizational Social Networks Using Minimum Spanning Tree," Proc. International Workshop On the Move to Meaningful Internet Systems (OTM 2008), Springer Press, Nov. 2008, pp. 18–19.

[32] J. Travers and S. Milgram, "An experimental study of the small world problem," Sociometry, 1969, pp. 425–443.

[33] S. Ye, J. Lang, and F. Wu, "Crawling online social graphs," Proc. IEEE International Asia-Pacific Web Conference (APWeb'10), IEEE Press, April. 2010, pp. 236–242.

[34] P. Zhao and C. Zhang, "A new clustering method and its application in social networks," Pattern Recognition Letters, vol. 32(15), 2011, pp. 2109–2118.

[35] Y. Zhou, K.R. Fleischmann, and W.A. Wallace, "Automatic text analysis of values in the enron email dataset: Clustering a social network using the value patterns of actors," Proc. IEEE Hawaii International Conference on System Sciences (HICSS 2010), IEEE Press, January 2010, pp. 1–10.