Analysis of On-Line Social Networks Represented as Graphs – Extraction of an Approximation of Community Structure Using Sampling

Néstor Martínez Arqué[1] and David F. Nettleton[1,2]

[1] Dept. Information Technology and Communications,
Universitat Pompeu Fabra, Barcelona, Spain
{nestor.martinez,david.nettleton}@upf.edu
[2] IIIA-CSIC, Bellaterra, Spain

Abstract. In this paper we benchmark two distinct algorithms for extracting community structure from social networks represented as graphs, considering how we can representatively sample an OSN graph while maintaining its community structure. We also evaluate the extraction algorithms' optimum value (modularity) for the number of communities using five well-known benchmarking datasets, two of which represent real online OSN data. Also we consider the assignment of the filtering and sampling criteria for each dataset. We find that the extraction algorithms work well for finding the major communities in the original and the sampled datasets. The quality of the results is measured using an NMI (Normalized Mutual Information) type metric to identify the grade of correspondence between the communities generated from the original data and those generated from the sampled data. We find that a representative sampling is possible which preserves the key community structures of an OSN graph, significantly reducing computational cost and also making the resulting graph structure easier to visualize. Finally, comparing the communities generated by each algorithm, we identify the grade of correspondence.

Keywords: Data mining, social networks.

1 Introduction

Finding structure in ad-hoc networks without any a priori knowledge about the expected result is a complex task. With the advent of online social networks (OSNs), the study of how to extract a vision of the network in terms of 'communities' has become an active field. By 'communities' we understand the 'sociological' interpretation in which individuals (humans beings) interact socially in some way (by email, using some online application, collaborating in some endeavour such as the writing of scientific papers, or forming some other social group such as a club, association, and so on). We can also extent our definition of individuals to include the study of the behaviour and social interactions of living beings in general (such as Dolphins, Simians and so on).

The results of extracting a community structure are highly dependent on the statistical and topological characteristics of the graph dataset, such as the average degree,

V. Torra et al. (Eds.): MDAI 2012, LNAI 7647, pp. 149–160, 2012.

clustering coefficient and level of fragmentation. It may also be dependent on the community extraction algorithm used, many of which are stochastic and non-deterministic. Other problems include the large volume of data in many OSN logs, the presence of 'noise' or 'unreliable' and outdated links in the graph, and highly fragmented graphs.

In the present study we apply two distinct community extraction algorithms [1,2] to five structurally distinct datasets, and compare the results. The extraction algorithms represent an optimization process based on an entropy type metric (modularity). For the three largest datasets we have also applied a filtering processing to reduce the number of nodes tested, while maintaining the key community structure information. This implies a very considerable saving in computational cost of processing by the community search algorithm. In this paper we describe how a filter based on degree and/or clustering coefficient enables us to extract the core parts of the communities in a complex graph. This filtering also significantly improves the results of visualization, using, for example, the Gephi software tool (http://gephi.org/), avoiding the typical "hairball" [3] appearance of many high data volume social networks.

The structure of the paper is as follows: in Section 2 we present the state of the art and related work; in Section 3 we present our approach for filtering and sampling the data; in Section 4 we define the datasets used and the experimental setup; in Section 5 we present the empirical tests and results for the community extraction, comparing the two algorithms and sampling Vs. using the complete dataset; finally in Section 6 we give the conclusions.

2 State of the Art and Related Work

The following briefly reviews the related work and key authors in the field of OSN graph processing, community detection and OSN graph sampling.

The study of community structure in social networks has been of interest for many years as a multidisciplinary field [4, 5]. 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, by authors such as [1, 6, 7, 8, 9], and which deal specifically with the mining of social networks as graphs [10, 11]. In this paper we benchmark two community structure extraction algorithms: Newman[1], which we have implemented in Python NetworkX and **(ii)** the Louvain method[2] using the default version available in the Gephi graph processing software.

Newman's algorithm[1] 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; (ii) identification of groups by their boundaries, using the least central edges (frontiers). This metric is also referred to as "edge betweenness", and is based on Freeman's "betweenness centrality measure" [5]. 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. Newman's fast algorithm [12] is used for calculating betweenness.

Newman's algorithm[1] extracts the communities by successively dividing the graph into components, using a metric to quantify the quality of the community partitions 'on the fly'. The value calculated by the quality metric for a given community is called the modularity. For a graph divided into k communities, a symmetrical matrix e of order $k2$ is defined whose elements e_{ij} are the subset of edges from the total graph which connect the nodes of communities i and j.

The modularity metric is defined as 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 [1], the usual empirical range for Q is between 0.3 and 0.7.

The Louvain method[2] 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 consists of 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.

With respect to the evaluation of community detection algorithms, Lancichinetti and Fortunato in [13] carried out an exhaustive benchmarking of 12 different methods, including the Louvain method[2] and Newman's method[1]. However, they used synthetic datasets for their tests and did not evaluate sampling of the networks. In the current work we have used real datasets and considered sampling. In [13], partition comparison between methods used the fraction of correctly identified nodes measure (NMI- Normalized Mutual Information). Lancichinetti also benchmarked a second measure, called 'LFR', which also takes into account degree power law distributions and community size. In our current work we have used Girvan and Newman's benchmark [14], using only the top N communities for evaluation, chosen by studying the size distributions.

Sampling is a 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 is related to, but not the same as filtering. Filtering eliminates records from the complete dataset according to some criteria, for example, "remove all nodes with degree equal to one". Sampling, on the other hand, 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. Chakrabarti, in [15], compares two sampling methods: *(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. In order to solve this problem, the authors propose a random or probabilistically weighted

selection of seeds. However, for community sampling, we propose that this bias is advantageous because we are interested in identifying key hubs and highly connected neighbors, as opposed to the more isolated regions and nodes of the graph. A key consideration in sampling is the choice of the initial starting nodes (or 'seeds') for extracting the sample. Another consideration is how to measure the 'quality' of the derived sample. These two aspects are studied in [16, 17].

3 Our Filtering/Sampling Approach

In this section, with reference to Figures 1a and 1b, we will see how we apply a two step process, consisting of filtering followed by sampling, in order to obtain a subset of a complete graph consisting of three communities. We emphasize that we have defined a process which is customized for extracting community structure. Thus we make emphasis on identifying hub nodes, high density regions, and their neighbors, rather than on an equitative percentage of all types of nodes. Hub nodes are identified by their degree, and high density regions are identified by the clustering coefficient. Once we have selected the "seed nodes" based on their degree or clustering coefficient, then we apply a sampling at 1 hop to obtain all the neighbors of each "seed node". Again, instead of applying a proportional number of nodes based on their distribution of the complete dataset, we let the search be biased to nodes with a high degree or a high clustering coefficient. In Fig. 1 we see a schematic representation of the filtering and sampling process.

Now we will see how we would process a simple graph consisting of 3 communities. In Fig. 2 we see the assigning of seeds (encircled nodes) using the 92.5 percentile of the degree *(a)* and clustering coefficient *(b)* values, respectively, and then including all the seed's neighbors (indicated by rectangles) at one hop. This means that the degree of the seed nodes (Fig. 2a) will be in the top 7.5% of the degree distribution for the complete graph. Likewise, the clustering coefficient of the seed nodes (Fig. 2b) will be in the top 7.5 of the distribution of the clustering coefficient for the complete graph.

We see that a very good coverage is obtained of the three communities in this graph, without having to expand the inclusion of nodes (in a "snowball" fashion) to 2 or more hops. This is because the regions we are interested in, the community cores, will be generally made up of a lattice of high degree and/or highly interlinked nodes. Therefore, selecting precisely these nodes as the seeds and including their neighbors will cover a high percentage of the core component of the major communities. In the empirical section we see how this result applies for much more complex and fragmented networks, and how we decide when to use the degree as the filter, or the clustering coefficient.

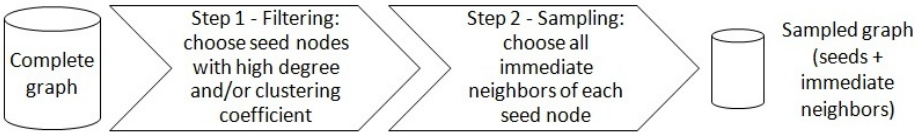
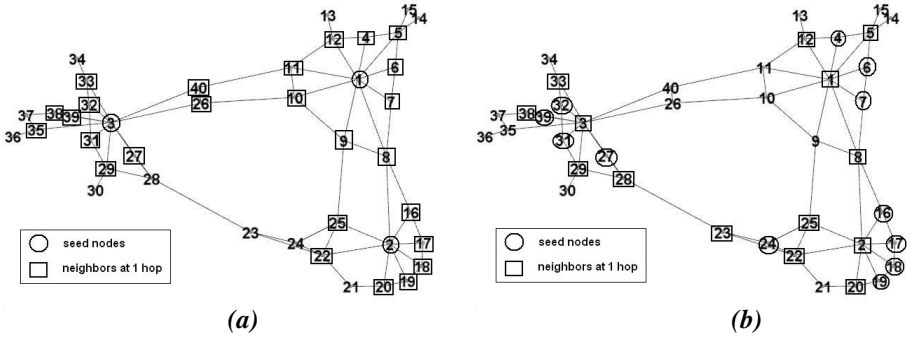


Fig. 1. Schematic representation of the node filtering and selection process



(a) *(b)*

Fig. 2. (a) Selection of seed nodes using 92,5 percentile of degree values**; (b)** Selection of seed nodes using 92,5 percentile of clustering coefficient values frequency

4 Datasets Used and Experimental Setup

In this section we briefly describe the procedure we have followed, the datasets used and their basic statistics. We also give the details of the sampling for the three large datasets, and their statistics after sampling. We use five different benchmark datasets: Karate [4], Dolphins [18], ArXiv GrQc-General Relativity and Quantum Cosmology [19], Enron [7] and Facebook [9]. In Table 1 we see a summary of the basic graph statistics for each test dataset. For community extraction we apply Newman's algorithm[1] (which we implemented in Python NetworkX) and the Louvain method[2] (standard version available in Gephi) to the datasets.

4.1 Values for Filtering and Sampling

With reference to Table 2, we see the summary for the sampling methods, per dataset. The filter and the value were chosen by different trial and error tests in order to obtain the desired overall percentage, which is the sum of the seed nodes plus all the neighbors of each of these. We used the recommendations of [20] as a guideline for the approximate optimum percentage of the complete dataset, which Ahn stated as being 25% for the degree as filter, and 20% for the clustering coefficient as filter. However, we found that the real sample size depends on the dataset and the distributions of the degree and clustering coefficient values.

Table 1. Summary of graph statistics for the five original datasets

	Karate	Dolphins	GrQc	Enron	Facebook
#Nodes	34	62	5242	10630	31720
#Edges	78	159	14496	164837	80592
Avg. degree	4.59	5.13	5.530	31.013	5.081
Clust. coef.	0.57	0.26	0.529	0.383	0.079
Avg. path length	2.408	3.356	6.049	3.160	6.432
Diameter	5	8	17	20	9

Table 2. Summary of sampling criteria/methods for sampled datasets

	Filter	Value	Resulting sample size	Sample
ArXiv-GrQc	Degree	≥30	17.91%	All neighbors
Enron	Clustering coef.	=1	20.83%	All neighbors
Facebook	Clustering coef.	≥0.5	10.75%	All neighbors

Table 3. Summary of graph statistics for the three largest sampled datasets

	GrQc	Enron	Facebook
#Nodes	939	2218	3410
#Edges	5715	14912	6561
Avg. degree	12.17	12.315	3.848
Clust. coef.	0.698	0.761	0.632
Avg. path length	4.51	3.143	8.388
Diameter	10	7	27

In Table 3 we summarize the basic graph statistics for each test dataset, after sampling. As we mentioned in Sec. 3, we are interested in extracting the strong community structure of the graph and therefore the fact that the before/after statistics are distinct is not an issue, which is caused mainly by the omission of isolated and low connectivity areas of the graph. One negative aspect would be the possible loss of some of the bridge nodes between communities, as commented in the previous section.

5 Empirical Tests and Results

In Section 5.1 we first document the results of applying Newman's method to the sampled datasets; then in Section 5.2 we compare the results with those of the literature; in Section 5.3 we apply the Louvain method and compare the communities extracted from the original datasets to those extracted from the sampled datasets, using an NMI type metric for node assignments to communities; finally, in Section 5.4 we compare the communities extracted by Newman's method and the Louvain method using the same NMI metric. N.B. *In the following text we will now refer throughout to Newman's method as NG and the Louvain method as LV.*

5.1 Evaluation of Newman's (NG) Method with the Sampled Datasets

For the **ArXiv-GrQc** dataset[19] and with reference to Fig. 3 (GrQc) and Table 4, the optimum modularity was obtained at Q=0.777, produced at iteration 56 and which partitioned the sampled version of the dataset in 57 communities. As can be seen in Fig. 3 (GrQc), the modularity value rises rapidly to a global maximum, which it maintains during approx. 100 iterations and then decays smoothly. With reference to Fig. 4a, the greatest community (lowest part of the Figure), is formed by 16.29% of the total nodes. The next two communities represent 10% of the total nodes.

For the **Enron** dataset, with reference to Fig. 3 (Enron) and Table 4, the optimum modularity was found at iteration 865, corresponding to Q=0.42, and dividing the dataset into 869 communities, of which the biggest represented 66.95% of the total nodes. We see from Fig. 2 that the modularity ascends rapidly during the first 70 iterations, and then keeps increasing with a much shallower gradient until it reaches the optimum, after which it begins to decay significantly. The version we finally used for the sampled Enron dataset was that generated an early iteration (51), which partitioned the dataset into 56 communities with a modularity close to the optimum obtained later at iteration 864.

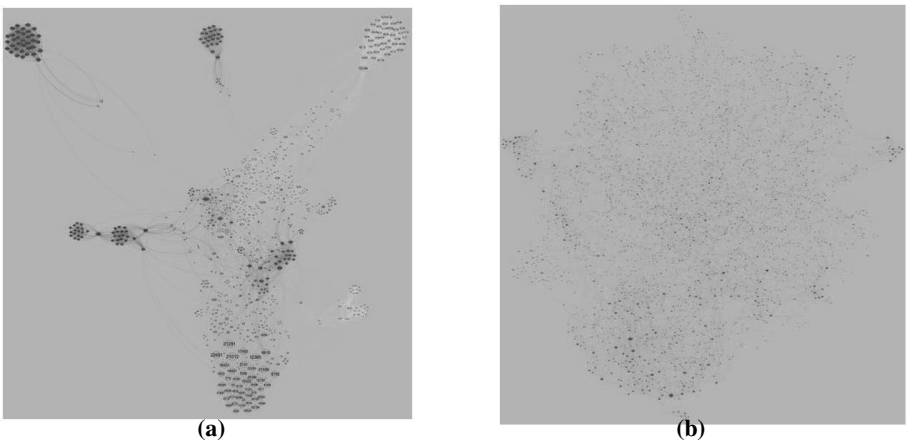
For the **Facebook** dataset, with reference to Fig. 3 (Facebook) and Table 4, the optimum modularity was found at iteration 40, with Q=0.87 (a relatively high value with respect to the other datasets), resulting in a total of 190 communities (Fig. 4b). This value was obtained as a consequence of the low clustering coefficient in the dataset. As can be seen in Fig. 3 (Facebook), the optimum value is found relatively early on in the process, followed by a linear decay from that point onwards.

Fig. 3. Evolution of the modularity Q for the three largest sampled datasets

Table 4. Summary of community structure processing statistics for five test datasets using the NG Method

	It.	Q	C	Original or Sampled
Karate	4	0.494	5	O
Dolphins	5	0.591	6	O
GrQc	56	0.777	57	S
Enron	865	0.421	869	S
Enron Early*	51	0.325	56	S
Facebook	40	0.870	190	S

It.=number of iterations, Q=modularity, C=number of communities, *Early termination with a semi-optimal Q.



(a) **(b)**

Fig. 4. NG **Method: (a)** Visualization of the principal communities extracted for the sampled arXiv-GrQc dataset; **(b)** Visualization for the Facebook dataset, with a partitioning corresponding to 190 communities obtained at iteration 40 (both using 'Force Atlas' visualization metric)

5.2 Discussion of Results with Reference to the Literature

In the following we reference the literature in order to obtain some idea of what we could consider the correct number of communities in each of the three largest datasets. For the arXiv-GrQc dataset, Xie[21] reported 499 communities using a "Label Propagation Algorithm" and 605 communities using a "Clique Propagation Algorithm". In the case of the Enron dataset, Shetty[7] defined the dataset as consisting of 151 Enron employees, and 5 key (hub) persons, however the graph is generated by emails sent between these persons, including "cc" and "re:" mailings to external emails, which greatly increases the number of entities in the dataset, and therefore nodes in the graph. Finally, for the Facebook dataset, there are no definitive values for the number of clusters in the literature, however Viswanath in [9] reported a high fragmentation into small communities, and Leskovec in [19] also reported a relatively high fragmentation of communities in other online social networks similar to Facebook. However the fragmentation of this particular dataset is also probably influenced by the measure of interaction (writes to wall) used to define links between users.

Table 5. Summary of community Q and C values for the three largest graphs using LV on the sampled and original versions of the datasets

	Original		**Sampled**	
	Q	**C**	**Q**	**C**
GrQc	0.856	390	0.789	11
Enron	0.491	43	0.560	68
Facebook	0.681	1105	0.519	33

Q=modularity, C=number of communities.

5.3 Communities Extracted using the LV Method from Sampled Datasets Vs. Communities Extracted using the LV Method from Original Datasets

In the following Section we have used the LV Method (rather than the NG Method) to compare the communities in the sampled and original datasets, given the very large computational cost of NG to process the larger (original) graphs, and because LV allows us to obtain an adequate benchmark for the datasets which is comparable to the results in the literature.

Table 5 shows a summary of the Q (modularity) and C (number of communities) values generated for the original and the sampled datasets. For the original datasets we see a high fragmentation of small communities in the arXiv-GrQc and Facebook datasets, which in the sampled datasets is greatly reduced so as to include only the most significant communities. Curiously, for the Enron dataset more communities are found in the sampled dataset than in the original dataset (68 with respect to 43), however for the former (sampled dataset) the number of communities it finds (value of 68 in Table 5) is similar to that of NG (value of 56 in Table 4).

NMI (Normalized Mutual Information). With reference to Table 6, we now compare the results of the community labeling by counting the number of nodes which are assigned in each community, and the number of nodes which are assigned to the same corresponding community, in the sampled and original datasets. This represents a NMI (Normalized Mutual Information) type metric [13] in which the nodes are labeled by the community of the original dataset, following the method we commented in Section 2. Table 6 summarizes the "purity" of the correspondences, in which a "purity" of 100% would mean that all the nodes which were assigned to communities $C_{1...}C_N$ in the sampled dataset were also assigned to communities $C_{1...}C_N$ in the original dataset.

The matching is made more difficult given that the LV Method (and the NG Method) are stochastic and non-deterministic. This means that each execution may produce slightly different results (nodes are assigned to different communities), although we assume the extraction of the most important communities will be similar. Also, the labels assigned to the communities (community 1, 2, etc.) may vary. Hence, in order to realize a comparison, we must first find the majority matching of each community label in the first execution with each community label in the second, in order to establish the correspondence. We chose the top N communities (in general N=10) by studying the size distribution.

In column B of Table 6 we see the difference for the sampled dataset with itself (for two different executions of the algorithm), and in column C we see the difference for the original dataset with itself. Hence, if we consider the correspondence of the

Table 6. Comparison of correspondence (NMI) of community assignments between original datasets and sampled datasets (using the LV Method)

	NMI orig. Vs. sampled (A)	NMI sampled Vs. sampled (B)	NMI orig. Vs. orig. (C)	Net loss (C - A)
GrQc	0.66559	0.82544	0.77301	0.10742
Enron	0.69069	0.86903	0.82012	0.12943
Facebook	0.58996	0.73249	0.69215	0.10219

original dataset with itself as the baseline (column C), then the net precision loss (last column of Table 6) will be the difference between the baseline and the correspondence between the communities of the original dataset with those of the sampled dataset (column A).

We observe from the final column of Table 6 that the precision loss is between 10% and 13%, depending on the dataset, and the average correspondence is between 58% and 70% (column A). The NMI of the sampled datasets (column B) represents a significant improvement with respect to the original datasets (column C). Finally, we inspected the correspondence between communities in the original dataset and those in the sampled dataset. We found that the most "pure" communities and the most "impure" in general remained the same, that is, communities with a high relative correspondence remained so and those with a low relative correspondence also remained so.

5.4 Communities Extracted by LV Method vs. Communities Extracted by NG Method

In Terms of 'Q' (Modularity Value) and 'C' (Number of Communities Created): We first compare the methods referring to Table 4 (NG, columns 3 and 4) and Table 5 (LV, 2 rightmost columns). In the case of the sampled version of the GrQc dataset, the number of communities extracted by LV (11) was different from that of NG (57). In terms of Q (modularity), both methods gave the same value (0.77 Vs. 0.79). For the sampled version of the Enron dataset, for NG we took the early cutoff version. The number of communities found was similar (56 for NG Vs 68 for LV), however the Q value was significantly lower for NG (0.32 for NG Vs. 0.56 for LV).

Finally, Facebook, gave the biggest difference in terms of C (190 communities for NG Vs. 33 for LV) and Q (0.87 for NG Vs. 0.52 for LV). We propose that a key factor in this result the lack of an identifiable cut-off point for NG, and the high fragmentation of communities in the Facebook dataset. In general, we can conclude that NG and LV may give distinct results in terms of the number of communities and modularity values.

In Terms of NMI (Normalized Mutual Information): In Table 7 we compare the assignments of nodes between the top N communities $\{C_{LV}\}$ extracted by LV and those extracted by NG $\{C_{NN}\}$, for the sampled data. We note that for column A we have used the N largest communities $\{C_{LV}\}$ created by LV, by number of nodes, then we find the percentage of corresponding nodes of the principal corresponding communities $\{C_{NN}\}$ of NG.

Table 7. Normalized Mutual Information (NMI) comparison of correspondence of node assignments to communities: LV Method Vs NG Method for sampled data

	NMI LV Vs. NG (A)	NMI NG Vs. LV (B)	NMI orig. Vs. orig. (C)	Net loss C - Avg. (A, B)
GrQc	0.69116	0.87243	0.77301	-0.00878
Enron	0.31313	0.68796	0.82012	0.31958
Enron early	0.83437	0.44320	0.82012	0.18133
Facebook	0.62056	0.54551	0.69215	0.10911

Contrastingly, in column B we first identify the top N communities $\{C_{NN}\}$ created by NG by number of nodes, then we find the percentage of corresponding nodes of the principal communities of LV $\{C_B\}$. In column C we define the same baseline used in Table 6, which is the NMI of the communities of the same dataset for two different executions of LV. Finally, in the final column we take the difference between the baseline and the average of columns A and B. From Table 7 we observe that the correspondence, in terms of node assignments, between the two methods is dataset dependent, with Enron (maximum number of iterations) having the least similarity (0.31) and GrQc having the greatest similarity (-0.01). We have also considered the 'early cutoff' version of applying NG to the Enron data, given that it produced a much smaller number of communities. As can be seen, there is a significant improvement with respect to the version which was allowed to run much longer (0.18 Vs. 0.31).

In conclusion with respect to the comparison of the methods, the empirical tests and results show there is a significant difference between the assignment of the nodes between methods.

6 Conclusions

We have benchmarked five statistically and topologically distinct datasets, applying two community structure elicitation algorithms and sampling on the three biggest datasets. The sampling is designed to maintain the overall community structure by choosing "hub" type nodes and high density regions, based on degree and clustering coefficient. The results indicate that it is possible to identify the principal communities for large complex datasets, using this type of sampling. The sampling method maintains the key facets of the community structure of a dataset, while reducing significantly (80 to 90%) the dataset size. We have also established, due to the stochastic nature of the algorithms, that a significant difference is found in the assignment of nodes to communities between different executions and methods. However, by inspection of the communities, we observe that the overall structure is consistent.

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

References

1. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 26113 (2004)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebure, E.: Fast unfolding of communities in large networks. In: J. of Stat. Mech.: Theory and Experiment (10), P1000 (2008)
3. Traud, A.L., Kelsic, E.D., Mucha, P.J., Porter, M.A.: Comparing Community Structure to Characteristics in Online Collegiate Social Networks. SIAM Review 53(3), 526–543 (2011)
4. Zachary, W.W.: An Information Flow Model for Conflict and Fission in Small Groups. Journal of Anthropological Research 33, 452–473 (1977)
5. Freeman, L.C.: A Set of Measures of Centrality Based on Betweenness. Sociometry 40(1), 35–41 (1977)

6. Newman, M.E.J.: Modularity and community structure in networks. PNAS 103(23), 8577–8582 (2006)
7. Shetty, J., Adibi, J.: Discovering Important Nodes through Graph Entropy - The Case of Enron Email Database. In: Proc. 3rd Int. W. on Link Discovery, pp. 74–81 (2005)
8. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattarcharjee, B.: Measurement and analysis of online social networks. In: Proc. 7th ACM SIGCOMM Conference on Internet Measurement, IMC 2007, pp. 29–42 (2007)
9. Viswanath, B., Mislove, A., Cha, M., Gummadi, K.P.: On the Evolution of User Interaction in Facebook. In: Proc. 2nd ACM Workshop on Online Social Networks, WOSN 2009, Barcelona, Spain, pp. 37–42 (2009)
10. Kleinberg, J.M.: Challenges in mining social network data: processes, privacy, and paradoxes. In: Proc. 13th Int. Conf. on K. Disc. & Data Mining (KDD 2007), pp. 4–5 (2007)
11. Kumar, R., Novak, J., Tomkins, A.: Structure and Evolution of Online Social Networks. In: Link Mining: Models, Algorithms, and Applications, Part 4, pp. 337-357. Springer (2010)
12. Newman, M.E.J.: Scientific collaboration networks. I. Network construction and fundamental results. Phys. Rev. E 64(1), 016131 (2001)
13. Lancichinetti, A., Fortunato, S.: Community Detection Algorithms: a comparative analysis. Physical Review E 80, 056117 (2009)
14. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)
15. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM Computing Surveys 38(1) (March 2006)
16. Bartz, K., Blitzstein, J., Liu, J.: Graphs, Bridges and Snowballs: Monte Carlo Maximum Likelihood for Exponential Random Graph Models. Presentation, January 8 (2009)
17. Lee, S.H., Kim, P.J., Jeong, H.: Statistical properties of sampled networks. Phys. Rev. E 73, 016102 (2006)
18. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology 54(4), 396–405 (2003)
19. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6(1), 29–123 (2009)
20. Ahn, Y.Y., Han, S., Kwak, H., Moon, S., Jeong, H.: Analysis of topological characteristics of huge online social networking services. In: Proc. 6th Int. Conf. WWW, pp. 835–844 (2007)
21. Xie, J., Szymanski, B.K., Liu, X.: SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-listener Interaction Dynamic Process. Cornell University Library (2011), `http://arxiv.org` arXiv:1109.5720v3