An Evaluation of Community Detection Algorithms on Large-Scale Email Traffic

Farnaz Moradi, Tomas Olovsson, and Philippas Tsigas

Computer Science and Engineering
Chalmers University of Technology, Göteborg, Sweden
{moradi,tomasol,tsigas}@chalmers.se

Abstract. Community detection algorithms are widely used to study the structural properties of real-world networks. In this paper, we experimentally evaluate the qualitative performance of several community detection algorithms using large-scale email networks. The email networks were generated from real email traffic and contain both legitimate email (ham) and unsolicited email (spam). We compare the quality of the algorithms with respect to a number of structural quality functions and a logical quality measure which assesses the ability of the algorithms to separate ham and spam emails by clustering them into distinct communities. Our study reveals that the algorithms that perform well with respect to structural quality, don't achieve high logical quality. We also show that the algorithms with similar structural quality also have similar logical quality regardless of their approach to clustering. Finally, we reveal that the algorithm that performs link community detection is more suitable for clustering email networks than the node-based approaches, and it creates more distinct communities of ham and spam edges.

Keywords: Community detection, Email networks, Quality functions

1 Introduction

Unfolding the communities in real networks is widely used to determine the structural properties of these networks. Community detection or clustering algorithms aim at finding groups of related nodes that are densely interconnected and have fewer connections with the rest of the network. These groups of nodes are called communities or clusters and they exist in a variety of different networks [9]. The problem of how to find communities in networks has been extensively studied and a substantial amount of work has been done on developing clustering algorithms (an overview can be found in [8, 21]). However, there is no consensus on which algorithm is more suitable for which type of network. Therefore, a number of studies have experimentally compared the qualitative performance of different community detection algorithms on synthetic and benchmark graphs with built-in community structure [12, 5]. However, these graphs are different from real-world networks as the assumptions they make are not completely realistic [8]. Delling et al. [6] have shown that the implicit dependencies between community detection algorithms, synthetic graph generators, and quality functions

used for assessing the qualitative performance of the algorithms make meaningful benchmarking very difficult. Therefore, empirical studies of the existing algorithms on real-world networks are crucial in order to evaluate different algorithms and to find the most suitable methods for different types of networks.

Moreover, community detection in real-networks has many different applications. Community detection algorithms can be used to find users with similar interests in a social network in order to provide recommendations to them, to group the peers that are geographically close in a peer-to-peer system to improve the performance of the system, or to detect the communities generated by malicious users in order to mitigate Sybil attacks [24]. In this paper, we study the community structure of a number of large *email networks* containing both legitimate *ham* and unsolicited *spam* emails. In an email network, the nodes represent email addresses and the edges represent email communications. In addition to a qualitative comparison of the algorithms, our goal is to find the best community detection algorithm that can separate spam and ham emails by clustering them into distinct communities. Such an algorithm can potentially be deployed in spam detection mechanisms that aim at mitigating the spam problem by looking at email traffic rather than email contents.

In order to achieve our goals, we have selected a number of broadly used community detection algorithms that are known to perform well on synthetic, benchmark, and a limited number of real graphs. In this study we evaluate and compare the qualitative performance of these algorithms when they are applied to large-scale email networks. Since the true community structure of our networks is unknown, it is important to use a quality measure to compare the algorithms. It is known that there is no single perfect quality metric for the comparison of the communities detected by different algorithms [2], therefore we use a number of *structural quality* functions such as modularity [17], coverage, and conductance [11], as well as a *logical quality* measure which is determined based on how homogeneous the edges inside the communities are. We use this measure to investigate and compare the ability of the selected algorithms in separating ham and spam emails into distinct communities.

The contributions of the paper are as follows. We show that there is a trade-off between creating high structural and high logical quality communities. Therefore, hierarchical and multiresolution algorithms which allow us to select the granularity of the clustering are more suitable to create the communities with the desired quality. We reveal that different algorithms that create communities with similar size distribution achieve similar structural and logical qualities, even though they use different approaches for clustering. Finally, we show that an algorithm that clusters networks based on the similarity of edges is superior to the algorithms that perform node-based clustering.

The rest of this paper is organized as follows. Section 2 presents the quality functions which are used for evaluating and comparing the algorithms. The community detection algorithms being compared are presented in Section 3. Section 4 reviews related previous research. In Section 5, the dataset used for empirical comparison is presented and the experimental results are discussed. Finally Section 6 concludes the work.

2 Quality of Community Detection Algorithms

In this section, we present the notations and the quality functions that are used in the rest of the paper.

Preliminaries Let $G = (V, E)$ represent a connected, undirected, and unweighted graph where V is the set of n nodes and E is the set of m edges of G. A *clustering* $\mathcal{C} = \{C_1, \ldots, C_k\}$ is a partitioning of V into k clusters C_i, by a node-based community detection algorithm. A cluster containing only a single node is called a *singleton*, and a cluster with only one internal edge is called *trivial*. If nodes can be shared between clusters, the clustering is called *overlapping*. The number of intra- and inter-cluster edges of a cluster C are denoted by $m(C)$ and $\overline{m}(C)$, respectively and $m(\mathcal{C}) := \sum_{C \in \mathcal{C}} m(C)$ is the total number of intra-cluster edges in $\mathcal{C}$.

Quality Functions A quality function is used either as an objective function to be optimized in order to find the communities of a network, or as a measure for assessing the quality of a clustering [6]. When the true community structure of a network is not known, quality functions are necessary for evaluating the qualitative performance of clustering algorithms. Since no single best quality function exists [2], we investigate three widely used structural quality functions: coverage, modularity [17], and conductance [11].

Coverage. Coverage of a clustering, $cov(\mathcal{C}) := \frac{m(\mathcal{C})}{m}$, is the most simple quality function, however, it is biased towards coarse-grained clusterings.

Modularity. Modularity is defined as $Q(\mathcal{C}) := \frac{m(\mathcal{C})}{m} - \frac{1}{4m^2} \sum_{C \in \mathcal{C}} \left(\sum_{v \in C} deg(v) \right)^2$ and is based on the idea that a good cluster should have higher internal and lower external density of edges compared to a *null model* with similar structural properties but without a community structure [17].

Conductance. Conductance of a cut $(C, V \backslash C)$ in a graph is defined as $\phi(C) := \frac{\overline{m}(C)}{\min(\sum_{v \in C} deg(v), \sum_{v \in V \backslash C} deg(v))}$, and tends to favor clusterings with fewer number of clusters [2]. Inter-cluster conductance, $\delta(\mathcal{C}) := 1 - \max_i \phi(C_i)$, $i \in \{1, \ldots, k\}$, is usually used as a worst-case measure to assess the quality of a clustering. The average conductance $(\frac{1}{|C|} \sum_{C \in \mathcal{C}} \phi(C))$ is also a useful metric, since if an algorithm creates singletons, the inter-cluster conductance value will be dominated by the zero value for these clusters, while the average would not [4].

The above widely used structural quality functions cannot be directly calculated for assessing the quality of link community detection methods because of the community overlaps. For instance, modularity of a link community can be calculated by applying a modified modularity function on a projected and weighted transformation of the network [7]. In this paper we investigate the structural quality of link communities by using two of the quality measures introduced in [1]. *Community coverage* measures the fraction of the nodes that belong to at least one non-trivial community, and *Overlap coverage* measures the average number of times a node is clustered inside non-trivial communities. Higher values for overlap coverage mean that the algorithm has extracted

more information from the network. The algorithms that don't find overlapping communities yield the same value for both overlap and community coverage.

In addition to the structural quality, we determine the *logical quality* of a clustering based on the type of the edges inside its communities. A clustering which yields only homogeneous communities, in which all of the edges are of the same type, has a perfect logical quality. For instance, a clustering with communities that contain only spam emails or only ham emails has higher logical quality compared to a clustering which yields communities containing a mixture of both ham and spam. In addition, the amount of spam and ham emails that can be separated into distinct homogeneous communities by an algorithm is used to determine its logical quality.

3 Studied Community Detection Algorithms

In this section we briefly describe the community detection algorithms we have selected and compared using our email networks.

Fast modularity optimization (Blondel) by Blondel et al. [3]. This algorithm, also known as *Louvain* method, is a greedy approach to modularity maximization. The algorithm starts with assigning each node to a singleton and progresses by moving nodes to neighboring clusters in order to improve modularity. This method has complexity $O(m)$ and unfolds a hierarchical community structure with increasing coarseness and meaningful intermediate communities.

Maps of random walks (Infomap) by Rosvall and Bergstrom [19]. This algorithm is a flow-based and information theoretic clustering approach with complexity $O(m)$. It uses a random walk as a proxy for information flow on a network and minimizes a *map equation*, which measures the description length of a random walker, over all the network clusters to reveal its community structure. Infomap aims at finding a clustering which generates the most compressed description length of the random walks on the network.

Multilevel compression of random walks (InfoH) by Rosvall and Bergstrom [20]. This method generalizes the Infomap method to reveal multiple levels of a network. InfoH minimizes a *hierarchical map equation* to find the shortest multilevel description length of a random walker.

Potts model community detection (RN) by Ronhovde and Nussinov [18]. This algorithm is based on minimization of the Hamiltonian of a local objective function, the absolute Potts model. The multiresolution variant of the algorithm deploys information theory-based measures to find the best communities on all scales. The complexity of this method is superlinear $O(m^{1.3})$ for the community detection algorithm and $O(m^{1.3} \log n)$ for the multiresolution algorithm.

Markov clustering (MCL) by Dongen [23]. MCL is based on the idea that a random walk entering a dense cluster likely remains for a long time inside the cluster before switching between sparsely connected communities. The random walks are calculated deterministically and simultaneously using a matrix of transition probabilities. The MCL algorithm has a complexity of $O(nk^2)$, where k refers to the average or maximum number of allowed neighbors for the nodes.

Link community detection (LC) by Ahn et al. [1]. All of the above algorithms aim at clustering nodes into densely connected communities. However, Ahn et

al. [1] have defined communities as a group of topologically similar edges and have introduced a link community detection algorithm for revealing them. The algorithm unfolds the hierarchical structure and overlapping communities of a network. Although the clustering is meaningful at all scales, an objective function, the *partition density*, is used to select the optimum level of hierarchy.

All of the above algorithms are known to perform well on large networks. Infomap, InfoH, and MCL are suitable for clustering networks where edges represent flows. Emails can be seen as flows of data between people, so flow-based approaches are good candidates for clustering email networks. Email communications can also be seen as pairwise relationships between people, so the other topological methods could also be suitable. LC which is based on calculating the similarity of the edges in a network can also be suitable since we are interested in grouping the same type of edges into the same clusters.

In this study, we have used the implementations of the algorithms, which were publicly available, in order to conduct the experiments. Blondel creates a hierarchy of clusterings where the best modularity is achieved at its last level. We have also looked at the clustering yield at Blondel's first level of hierarchy, which has smaller meaningful communities, and refer to it as *Blondel L1*. We have also used the basic RN algorithm instead of its multiresolution variant to be able to choose the desired clustering granularity. The granularity of the clusterings should be considered when comparing the quality of the algorithms since structural quality functions are usually in favor of coarse-grained clusterings [2].

4 Related Work

Experimental comparisons of different community detection algorithms have been conducted on both real and benchmark graphs. Lancichinetti and Fortunato [12] compared different algorithms including Blondel, Infomap, RN, and MCL, on GN and LFR benchmark graphs. They showed that Infomap, Blondel, and RN perform well, but MCL performs worse especially for large communities. They also showed that the performance of Blondel decreases for large graphs, whereas Infomap remains stable. Brandes et al. [4] conducted an experimental evaluation of three clustering methods including MCL using random clustered graphs and showed that MCL performs well with respect to some quality functions but produces more clusters than contained in the network.

Community detection algorithms have also been evaluated and compared using different real networks. Tibély et al. [22] have analyzed the community structure of a large mobile phone call graph using Blondel L1, Infomap, and an overlapping method. Leskovec et al. [14] studied a number of real networks, including the Enron email network and an email network of a large organization, to empirically compare two different clustering methods. The latter dataset was also used by Lancichinetti et al. [13], in addition to other real networks, to study the characteristics of communities in different types of complex networks. They used Infomap together with another algorithm to show that although different methods output different clusterings, the statistical properties of their communities are quite similar for similar classes of networks. Studies of the community

structure of email networks have also been conducted by Guimerà et al. [10] using emails in a university.

In contrast to previous studies, the dataset used in this study is based on email traffic captured on a high speed Internet backbone link, and is not limited to a single organization. To the best of our knowledge, this is the first study of the community structure of large-scale email networks containing spam. This dataset enables us to evaluate the ability of the community detection algorithms in separating spam from legitimate email by clustering them into distinct clusters.

5 Experimental Evaluation

In this section, the email dataset and the the experimental results are presented.

5.1 Dataset

The dataset used for creating the email networks was generated by collecting SMTP packets on a 10 Gbps link of the core-backbone of SUNET[1] during a period of 14 consecutive days in March 2010. During the collection period more than 797 million SMTP packets were collected, which were sent and received by 614,601 distinct domains. Around 3.4 million emails were extracted from the collected packets after removing unusable and rejected email transmissions. These emails were then classified to be either *spam* or *ham* using a well-trained filtering tool [2]. Following that, email contents were discarded and email addresses were anonymized in order to preserve privacy in a way that no information about the senders, receivers, and content of the emails are retrievable.

In addition to a complete email network, we generated daily and weekly email networks. An email network consists of email addresses as nodes, and the email communications between them as edges. More details on the measurement location, data collection and pre-processing, and the structural and temporal properties of the email networks can be found in [15] and [16], respectively.

5.2 Comparison of the Algorithms

In this section, the experimental results regarding the qualitative performance of the clustering algorithms with respect to their structural and logical quality is presented. A summary of the results can be found at the end of the section.

Table 1 shows the total number of nodes and edges, and the number of spam edges in each studied email network, as well as the number of communities created by each clustering algorithm. The algorithms were applied to the giant connected component (GCC) of each email network, which is a subset of the nodes in the network where a path exists between any pair of them. The networks are also considered as unweighted and undirected.

[1] The Swedish University Network (`http://www.sunet.se/`) serves as a backbone for university traffic, student dormitories, research institutes, etc.

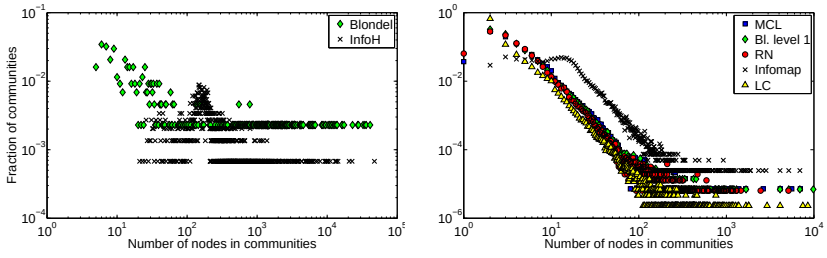
[2] The SpamAssassin (`http://spamassassin.apache.org`) was in use for a long time in our University mail server and it incurs high detection and low false positive rates.

Table 1. The properties of the GCC of the generated email networks (larger networks become more connected) and the number of communities created by each algorithm.

Network	# Nodes	# Edges	# Spam	Blondel	InfoH	Infomap	Blondel L1	MCL	RN	LC
Day 1	167,329	236,673	173,640	253	546	10,505	39,477	38,775	41,215	88,028
Day 2	153,734	194,797	97,260	194	397	8,025	28,077	27,011	28,499	61,027
Day 3	123,878	168,896	108,996	218	412	8,151	29,150	28,031	30,022	64,310
Day 4	128,200	172,836	113,299	218	398	8,484	29,123	28,043	30,167	63,165
Day 5	101,643	135,195	89,119	195	311	6,664	22,212	21,593	23,935	46,928
Day 6	72,068	99,361	75,713	236	183	4,714	13,904	13,716	17,697	30,236
Day 7	73,131	103,293	85,879	199	271	4,842	17,305	16,808	18,631	37,581
Week 1	901,699	1,441,731	961,809	558	1,470	41,916	149,131	144,054	187,960	451,275
Day 8	115,232	155,919	90,299	234	379	7,745	27,661	26,514	28,409	57,931
Day 9	112,713	152,569	88,273	188	383	7,521	26,395	25,549	26,942	56,443
Day 10	140,843	195,999	121,158	255	441	8,664	31,033	30,231	39,020	67,741
Day 11	125,029	179,410	116,056	192	398	8,171	28,501	27,897	30,484	65,285
Day 12	106,816	149,407	100,595	211	380	7,319	25,314	24,328	28,040	54,317
Day 13	73,325	98,713	71,954	339	296	5,275	16,736	16,074	22,476	32,403
Day 14	68,315	100,089	76,408	179	210	4,741	14,567	14,254	17,822	31,463
Week 2	810,543	1,348,373	859,324	436	380	40,553	143,569	139,366	156,822	430,232
All	1,599,732	2,790,322	1,858,686	1,028	1,740	63,471	230,013	220,346	294,581	817,074

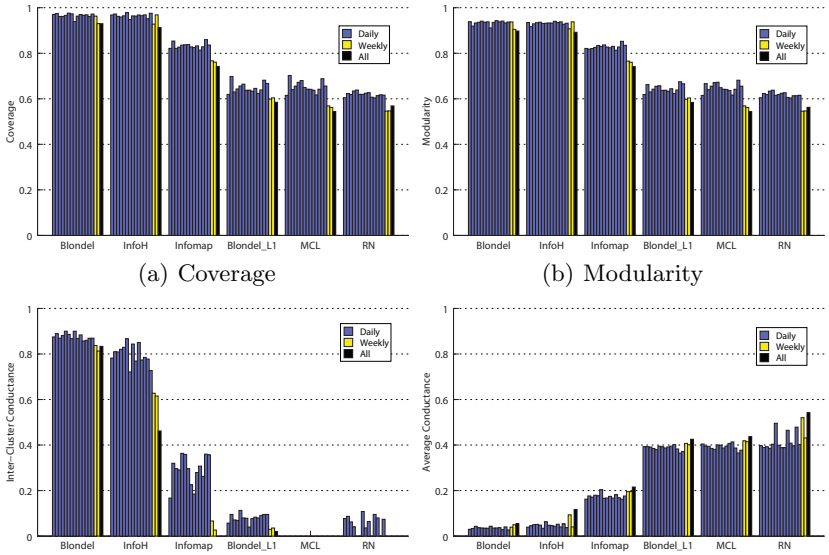
Blondel creates a coarse-grained clustering and in average achieves 46% modularity gain over Blondel L1. InfoH also creates coarse clusters and in average gains more than 15% in the compression of the description length of the random walks on the networks over the non-hierarchical version (Infomap). MCL allows us to select the granularity of the clustering by choosing an inflation parameter. It is also possible to choose the resolution parameter for RN to achieve a clustering with the desired granularity. We have selected the inflation parameter in MCL and the resolution parameter in RN so that for most of the networks they yield clusterings with a close granularity to that of Blondel L1. This allows us to further investigate and compare the effect of the granularity of the clusterings on their quality. LC is different in nature from the other algorithms as it is based on link community detection rather than a node-based approach. LC yields the finest-grained clustering for all of the networks at its best level of hierarchy.

Figure 1 summarizes the distribution of the size of the communities created by the different algorithms for the "week 2" email network. The distributions for other daily and weekly networks are quite similar. It can be seen that Blondel and InfoH, which create very coarse-grained clusters, have very different community size distributions compared to each other and the rest of the algorithms. It can



(a) (b)

Fig. 1. A comparison of community size distribution using "Week 2" email network. Blondel L1, MCL, and RN follow very similar distributions.



(a) Coverage (b) Modularity

(c) Inter-cluster conductance (d) Average conductance

Fig. 2. Comparison of structural quality of the algorithms on daily, weekly, and complete email networks. Blondel and InfoH yield the best structural quality.

also be seen in Figure 1(b) that, surprisingly, Blondel L1, MCL, and RN follow similar distributions. The main difference is that MCL and RN create a number of singletons, but Blondel L1 does not. The community size distribution of LC is also close to the other three methods, but it creates more clusters.

Structural Quality Figure 2 shows a comparison of the structural quality of the different clusterings. Each bar corresponds to a daily network (day 1 to day 14), except the last three bars from the left for each of the algorithms, which correspond to week 1, week 2, and complete email networks, respectively. It can be seen that Blondel, which aims at maximizing modularity, have the highest structural quality with respect to all of the quality functions. Although InfoH uses a fundamentally different approach it achieves equally good structural quality, however its quality degrades for larger networks. Blondel L1, MCL, and RN, which have closer granularities, also show similar quality with respect to coverage, modularity, and average conductance. However, based on the inter-cluster conductance, MCL and RN do not perform well since they might create a number of singletons which results in an inter-cluster conductance of zero.

Our experimental results reveal that the structural quality of clusterings are roughly consistent for different daily networks. The clusterings with similar granularity and community size distribution also show similar structural quality, therefore, it is important to take the granularity of the clusterings into account when comparing the algorithms. LC creates a clustering with the finest granularity, however the studied structural quality functions cannot be directly used for

assessing the quality of this algorithm due to its different nature. In this paper, we look at community coverage and overlap coverage which were introduced for assessing the quality of link-based clustering by Ahn et al.[1].

LC, Blondel, and InfoH yield full community coverage for all of the email networks. Infomap, Blondel Ll, MCL, and RN achieve community coverage of around 0.99, 0.84, 0.83, and 0.8, respectively. However, this function on its own is not enough for assessing the quality of a clustering method, it is also important to consider the overlap coverage of the clusterings. None of the algorithms, except MCL and LC, find overlapping clusters so their overlap coverage is equal to their community coverage. MCL is not an overlapping clustering method, but for some specific graphs it might find overlaps [23]. In our email networks, MCL yields very few overlapping communities so its overlap coverage is just slightly higher than its community coverage. LC yields overlap coverage of 2.6, 3.1, and 3.4 in average for the daily, weekly, and complete email networks, meaning that it unfolds more overlaps in larger networks.

Logical Quality Our experiments show that all algorithms create a number of *spam communities* that only contain spam, *ham communities* that only contain ham, and *mix communities* with a mixture of both ham and spam edges. Figure 3 shows a comparison between the percentage of spam, ham, and mix communities created by the different algorithms. The last three bars from the left for each of the algorithms correspond to week 1, week 2, and the complete email networks, respectively. It can be seen that InfoH and Blondel perform worse, since these algorithms tend to merge smaller homogeneous communities into mix communities to achieve higher structural quality. The best results for all networks are achieved by LC.

Moreover, it is important to assess the amount of spam and ham emails that can be separated by community detection algorithms, in order to investigate the possibility of deploying clustering approaches to perform spam detection. Figure 4 shows the ratio of total spam and ham edges that are inside homogeneous spam and ham communities. In all of the networks, communities created by LC contain the highest number of spam and ham edges. Blondel and InfoH have the worst logical quality and Blondel L1, MCL, and RN have almost similar quality. For all algorithms, except LC, some of the spam and ham emails end up as inter-cluster edges and can therefore not be separated by the clustering algorithms. It can also be seen that the percentage of spam (ham) edges which are clustered inside spam (ham) communities decreases for larger networks.

Our experiments suggest that the logical quality tends to be higher for fine-grained clusterings. The granularity of the best clustering created by LC is finer than the other clusterings in our experiments. LC cuts its hierarchy of clustering at an optimum threshold which results in maximal partition density. By choosing a threshold below the optimum value, we can have a clustering with coarser granularity. Since the algorithm reveals meaningful communities at all scales, we changed the threshold so that the granularity of the clustering became more similar to that of Blondel L1, MCL, and RN. Our experiments with the new clusterings show that, the percentage of spam (ham) edges inside the spam (ham) communities was reduced. For instance, for the first daily network the percentage

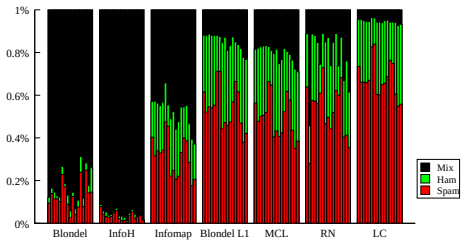
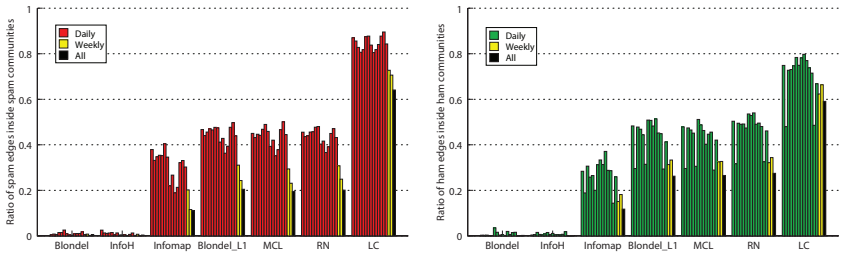


Fig. 3. Comparison of percentage of spam, ham, and mix communities created by different algorithms. LC creates the highest number of homogeneous communities.



(a) Spam communities (b) Ham communities

Fig. 4. Ratio of spam (ham) in homogeneous spam (ham) communities. LC clusters a higher percentage of total spam (ham) edges inside the spam (ham) communities.

of spam (ham) edges decreased from 87% to 66% (from 76% to 56%). Although the logical quality degrades by changing the coarseness of the clustering, LC still shows higher logical quality than all of the other algorithms.

Summary of the Experimental Results

- Blondel and InfoH create coarse-grained clusters and achieve the best quality with respect to all of the structural quality functions. However, they have the worst logical quality with respect to both number of homogeneous communities and amount of spam and ham emails that are clustered inside these homogeneous communities.
- Infomap, which is the non-hierarchical version of InfoH, achieves quite good structural quality and decent logical quality. However, Blondel L1, which is based on the first level of Blondel's hierarchy of clusterings, yields much better logical quality than Infomap, but worse structural quality with respect to all of the structural quality functions.
- MCL and RN allow us to change the resolution of the clustering by modifying different parameters. When the granularity of their clusterings is set to be close to that of Blondel L1, they show almost similar community size distribution as well as similar structural and logical quality. However, Blondel L1 is superior to the other two methods due to its lower complexity.

– LC, which performs link community detection, has the best logical quality and separates the highest amount of spam and ham emails into distinct homogeneous communities.

6 Conclusions

In this study, we have performed an empirical comparison and evaluation of a number of high quality community detection algorithms using large-scale email networks. The studied email networks contain both legitimate and spam emails and are created from real email traffic. Our study reveals that yielding high structural quality by community detection algorithms is not enough to unfold the true logical communities of the email networks. Therefore, it is necessary to deploy more realistic measures for clustering real-world networks.

More specifically, our study suggests that the community detection algorithms that achieve maximum modularity, coverage, inter-cluster conductance, or minimum average conductance do not reveal the communities that coincide with the true clustering of the email networks. For instance the algorithms which yield worse, but acceptable, average conductance values actually could separate a large number of spam (ham) emails into distinct spam (ham) communities. Therefore, the value of this function can be indicative of good logical quality. However, this observation is based on our email networks, and might not be conclusive as it was shown that different classes of networks show different community structures [12, 2].

Overall, our experiments reveal that link community detection is the most suitable approach for separating spam and ham emails into distinct communities compared to the other node-based algorithms.

Acknowledgments This work was supported by .SE – The Internet Infrastructure Foundation and SUNET. The research leading to these results has also received funding from the European Union Seventh Framework Programme (FP7/ 2007-2013) under grant agreement no. 257007.

References

1. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–4, Aug. 2010.
2. H. Almeida, D. Guedes, W. Meira Jr., and M. J. Zaki. Is There a Best Quality Metric for Graph Clusters? In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, pages 44–59. Springer-Verlag, 2011.
3. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct. 2008.
4. U. Brandes, M. Gaertler, and D. Wagner. Experiments on Graph Clustering Algorithms. In *Proceedings of the 11th European Symposium on Algorithms*, pages 568–579. Springer-Verlag, 2003.

5. L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008–P09008, Sept. 2005.
6. D. Delling, M. Gaertler, G. Robert, Z. Nikoloski, and D. Wagner. How to Evaluate Clustering Techniques. Technical report, no. 2006-4, Universität Karlsruhe, 2006.
7. T. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1):1–8, July 2009.
8. S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, Feb. 2010.
9. 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):7821–6, June 2002.
10. R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 68(6 Pt 2):065103, Dec. 2003.
11. R. Kannan, S. Vempala, and A. Veta. On clusterings-good, bad and spectral. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 367–377. IEEE Comput. Soc, 2000.
12. A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):1–11, Nov. 2009.
13. A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato. Characterizing the community structure of complex networks. *PloS one*, 5(8):e11976, Jan. 2010.
14. J. Leskovec, K. J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, page 631, New York, New York, USA, 2010. ACM Press.
15. F. Moradi, M. Almgren, W. John, T. Olovsson, and P. Tsigas. On Collection of Large-Scale Multi-Purpose Datasets on Internet Backbone Links. In *Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, 2011.
16. F. Moradi, T. Olovsson, and P. Tsigas. Structural and Temporal Properties of E-mail and Spam Networks. Technical report, no. 2011-18, Chalmers University of Technology, 2011.
17. M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):1–15, Feb. 2004.
18. P. Ronhovde and Z. Nussinov. Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, 80(1):1–18, July 2009.
19. M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–23, Jan. 2008.
20. M. Rosvall and C. T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, Jan. 2011.
21. S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, Aug. 2007.
22. G. Tibély, L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki. Communities and beyond: Mesoscopic analysis of a large social network with complementary methods. *Physical Review E*, 83(5):1–10, May 2011.
23. S. VAN Dongen. *Graph clustering by flow simulation*. PhD thesis, University of Utrecht, The Netherlands, 2000.
24. B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove. An analysis of social network-based Sybil defenses. In *Proceedings of the ACM SIGCOMM 2010 conference*, page 363, New York, New York, USA, 2010. ACM Press.