Tracking Communities of Spammers by Evolutionary Clustering

Kevin S. Xu XUKEVIN@UMICH.EDU
EECS Department, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122 USA

Mark Kliger MARK@MEDASENSE.COM
Medasense Biometrics Ltd., PO Box 633, Ofakim, 87516 Israel

Alfred O. Hero III HERO@UMICH.EDU
EECS Department, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122 USA

Abstract

We consider the problem of tracking communities in social networks over time, which is a natural extension of community detection to dynamic networks. The network of interest in this study consists of interactions between email spammers inferred by common usage of resources. We perform evolutionary spectral clustering on this network to reveal communities of spammers and track how they change with time. We conclude with a discussion of open problems and challenges presented by community tracking.

1. Introduction

A fundamental problem in the analysis of social networks is the detection of communities. A community is often defined as a group of nodes with much stronger ties to nodes within the group than to nodes outside of the group. Most studies of community detection in social networks have focused on static networks. The presence of the World Wide Web has given researchers access to dynamic social network data that is typically captured over a large period of time. Hence, the natural extension of community detection to dynamic networks is community tracking, which makes it possible to observe how communities grow, shrink, merge, or split with time.

In this study we track communities in a network of interactions between email spammers. Unlike with most social network data where interactions between nodes are directly observed, the interactions in this network must be inferred from resource sharing. The data source for this study is Project Honey Pot[1] and is described in Section 2.

We make use of evolutionary spectral clustering to track communities of spammers over time. Spectral clustering is a popular method for community detection in static networks, and evolutionary spectral clustering provides us with the necessary extension to dynamic networks. In short, it assigns nodes to communities at each time so that the communities are representative of both current and past data. The evolutionary spectral clustering procedure and its application to tracking communities are discussed in Section 3. Finally, we present some preliminary analysis results along with a discussion of open problems and challenges in community tracking in Sections 4 and 5.

2. Project Honey Pot data set

Project Honey Pot is a distributed system for monitoring harvesting and spamming activity via a network of decoy web pages with trap email addresses embedded within the HTML source, known as honey pots. Harvesting is the process by which spammers acquire email addresses. Project Honey Pot provides us with the IP address of the harvester and email server used for each spam email received at a trap address. A previous study on the Project Honey Pot data (Prince et al., 2005) found that while spammers often send spam in a distributed manner (e.g. by using botnets), harvesting is typically done in a centralized manner. Thus harvesters are likely to be associated with spammers, and in this study we assume that the harvesters monitored by Project Honey Pot are indeed representative of spammers. This allows us to associate each spam email with a spammer so that we can track communities of spammers. We shall refer to harvesters as spammers in the following, with the understanding that they are not the same entities but are very closely associated.

Project Honey Pot has grown exponentially over time, with over 85,000 spammers tracked as of March 2010. The number of active spammers monitored by Project Honey Pot during 2006 and 2007 is shown in Figure 1. We focus our attention on the data collected during 2006, prior to the rapid growth. We divide the data set into equal time intervals, which we denote by the superscript t. At each time t, the data can be represented by an $m^t \times n^t$ matrix H^t, where m^t denotes the number of active spammers at time t, and n^t denotes the number of

[1] Additional information on Project Honey Pot is available at the web site `http://www.projecthoneypot.org`.

servers used. Each entry of H^t is given by $h_{ij}^t = p_{ij}^t / e_i^t$, where p_{ij}^t denotes the number of emails sent by spammer i using server j during time interval t, and e_i^t denotes the number of addresses acquired by spammer i from the initial time interval up to time t. e_i^t is a normalization term to remove observation bias due to spammers acquiring different numbers of addresses and allows us to characterize behavior on a per-address basis.

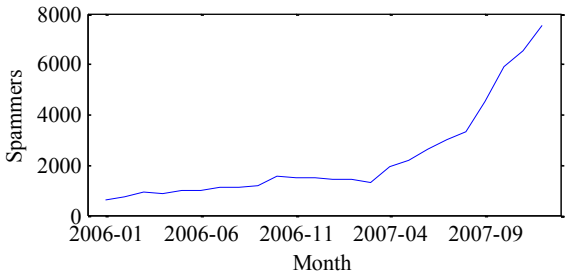

Figure 1. Number of active spammers in each month tracked by Project Honey Pot.

3. Methodology

The objective of this study is to track communities of spammers over time. Since we are unable to observe direct interactions between spammers, we use indirect interactions to infer their social network. In particular, we take the ties between spammers to correspond to common usage of resources, namely the servers used to send spam emails. We represent the social network of spammers by a dynamic graph $G^t = (V^t, E^t, W^t)$ where V^t is the set of nodes at time t, representing active spammers, E^t is the set of edges between nodes, representing inferred social ties between spammers, and W^t is the matrix of edge weights denoting the strengths of the ties, commonly referred to as the affinity matrix. Since h_{ij}^t denotes the normalized number of emails sent by spammer i using server j during time interval t, we take the weight of the edge between spammers i_1 and i_2 to be the dot product of rows i_1 and i_2 of H^t.

The problem of identifying communities in the social network translates into finding a graph partition that maximizes similarity between nodes in the same group and minimizes similarity between nodes in different groups. Spectral clustering (von Luxburg, 2007) is a commonly used method for community detection in static graphs.

3.1 Evolutionary spectral clustering

A method for evolutionary spectral clustering has recently been proposed (Chi et al., 2007) to extend spectral clustering to dynamic data. First we note that is possible to perform spectral clustering on dynamic data simply by clustering at each time using the most recent data. This approach has two main disadvantages: it is extremely sensitive to noise, and it also produces clustering results that are inconsistent with results from previous time intervals.

The objective of evolutionary spectral clustering is to identify communities that are representative of both current and past data, resulting in clustering results that are smooth over time. Chi et al. (2007) showed that this could be accomplished by performing ordinary spectral clustering on a convex combination of current and past affinity matrices. Furthermore, a method for estimating the optimal weights in the convex combination was proposed in (Xu et al., 2010), which we summarize in the following.

We define the smoothed affinity matrix at time t to be $\bar{W}^t = \alpha^t \bar{W}^{t-1} + (1-\alpha^t) W^t$ for $t \geq 1$ and $\bar{W}^0 = W^0$. α^t can be interpreted as a forgetting factor that controls the amount of smoothing to be applied. It was found that the forgetting factor that minimizes the squared Frobenius error $\mathrm{E}[\|\bar{W}^t - \mathrm{E}(W^t)\|_F^2]$ is given by

$$\alpha^t = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathrm{var}\left(w_{ij}^t\right)}{\sum_{i=1}^n \sum_{j=1}^n \left\{\left[\bar{w}_{ij}^{t-1} - \mathrm{E}\left(w_{ij}^t\right)\right]^2 + \mathrm{var}\left(w_{ij}^t\right)\right\}}.$$

In a real application however, the means and variances of the entries of W^t are not known, so the optimal α^t cannot be computed. It can, however, be approximated by replacing the unknown means and variances with sample means and variances. We refer interested readers to (Xu et al., 2010) for the details of the implementation.

3.2 Tracking communities over time

There are several additional challenges in order to track communities over time. While evolutionary spectral clustering provides a clustering result at each time that is consistent with past results, one still faces the challenge of matching communities at time t with communities at time $t-1$. Also, communities can merge or split, and nodes may enter or leave the network at various times.

We address the first issue by performing majority vote between community memberships at time t and $t-1$ to match communities between time intervals. We first match the largest community at time t, then the second largest, and continue in this fashion until all communities have been exhausted.

The issue of communities merging or splitting is intimately related to the problem of identifying the number of communities in a graph. Any heuristic for choosing the number of communities in ordinary spectral clustering, such as the eigengap heuristic (von Luxburg, 2007), can also be used in evolutionary spectral clustering by applying it to $\bar{W}^t$ instead of W^t.

Finally, the issue of nodes entering or leaving the network can be dealt with in the following manner. Nodes that leave the network between times $t-1$ and t can simply be removed from $\bar{W}^{t-1}$. Nodes entering the network at time t do not have corresponding rows or columns in $\bar{W}^{t-1}$ to perform smoothing with. These nodes can be removed from W^t in order to estimate α^t, then re-inserted into $\bar{W}^t$ so that they will be clustered as well.

4. Preliminary results

We applied evolutionary spectral clustering to the Project Honey Pot data from 2006 at one-month time intervals. The estimated forgetting factor α^t at each month is shown in Figure 2.

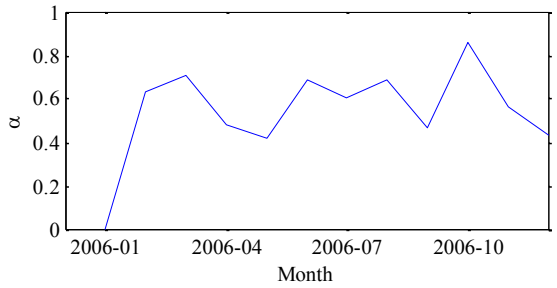

Figure 2. Estimated forgetting factor α^t at each month.

Notice that α^t drops around April, September, and December, suggesting changes in the community structure during these months. The community memberships by month are displayed in Figure 3 for the 240 spammers who were active for the entire year. Indeed in April, September and December, there are significant changes in the community structure. A large change also takes place in July, but there is only a small decrease in α^t for July.

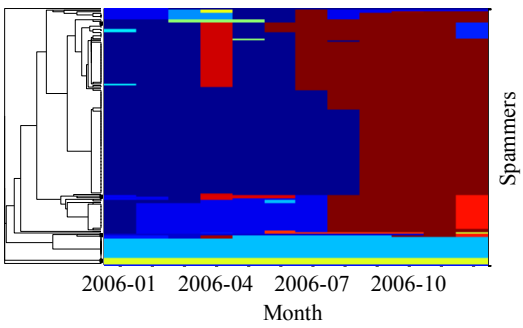

Figure 3. Heat map of community memberships by month. Each row of the heat map corresponds to a spammer, and each color denotes a community. The rows have been grouped by average linkage hierarchical clustering as indicated by the dendrogram on the left of the heat map.

One particular challenge in our analysis is the selection of the number of communities at each time. We applied the eigengap heuristic on the smoothed affinity matrix $\bar{W}^t$, but many gaps between eigenvalues appear, so it is not a definitive criterion for selecting the number of communities. A poor choice for the number of communities may create the appearance of communities merging or splitting when there is no actual change occurring.

5. Discussion

Our analysis has highlighted several major challenges that temporal tracking of communities presents in addition to the challenges present in static community detection. Validating a clustering result is already a difficult task in ordinary clustering, especially if one deals with an unlabeled social network. For example in (Xu et al., 2009) we validate communities of spammers by demonstrating that communities revealed by ordinary spectral clustering divide into communities of phishing spammers and communities of non-phishing spammers. Evolutionary clustering also adds the challenge of validating changes in communities over time. One possible method for doing so is to compare α^t with times of known major events or change points, if this information is available. Further research is required to develop additional validation techniques for evolutionary clustering results.

Another major challenge and an open problem in both static and evolutionary clustering is the selection of the number of communities. The availability of data at multiple time intervals may actually simplify this problem since one would expect that the number of communities, much like the community memberships, should vary smoothly with time. Hence, the development of methods specifically for selecting the number of clusters in evolutionary clustering is another interesting area of future research.

Acknowledgments

The authors would like to thank Unspam Technologies Inc. for providing us with the Project Honey Pot data. This work was partially supported by NSF grant CCF 0830490 and ONR grant N00014-08-1-1065.

References

Chi, Y., Song, X., Zhou, D., Hino, K., and Tseng, B. L. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.

Prince, M., Holloway, L., Langheinrich, E., Dahl, B. M., and Keller, A. M. Understanding how spammers steal your e-mail address: An analysis of the first six months of data from Project Honey Pot. In *Proceedings of the 2nd Conference on Email and Anti-Spam*, 2005.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–415, 2007.

Xu, K. S., Kliger, M., Chen, Y., Woolf, P. J., and Hero III, A. O. Revealing social networks of spammers through spectral clustering. In *Proceedings of the IEEE International Conference on Communications*, 2009.

Xu, K. S., Kliger, M., and Hero III, A. O. Evolutionary spectral clustering with adaptive forgetting factor. In *Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010.