Project Proposal

Sensing Semantic Information from Mobile Social Networks

Ting Zhang and Wen Yi
School of Industiral Engineering, School of Electrical and Computer Engineering
zhan1013@purdue.edu, yi35@purdue.edu

October 1, 2015

1 Introduction

Human society consists of extensive communications and interactions between individuals, via the use of mobile sensors, such as mobile phones, tablets and GPS. The understanding of individual relations from these sensors, can greatly facilitate and promote the interactions between individuals. For example, listing phone contacts in semantic orders according to the time and location when a person wants to make a phone call, would save the person both time and memory load, to find a specific contact from a phone book with tons of contacts based on alphabetic order. Further more, by understanding the friendship network from the global view, we can group people in the whole network into small communities in which people have closer friendships. This may serve as a friend grouping suggestion function inside the cellphone contact managing software. Therefore, in this project, we will focus on the inference of friend relationships with the detection of communities using data collected from mobile phones. Detailed explanation of dataset is stated in section 4.

2 Related work

2.1 Relationship Inference

Relationship inference in social networks has been studied in various fields and domains. In this context, we refer to friendship inference between pairs of individuals. Representing social networks with topology structures provides insights to predict relationships between individuals based on topology and probability distribution of the links in the topology. Liben-Nowell and Kleinberg proposed different measurements to compute the "similarity" between two nodes (individuals) in the graph, including the distance between two nodes, number of shared neighbors, and "meta-approaches" that integrate different measurements. Beyond topological structures, individual attribute and context information have also been utilized to facilitate the construction of relations between individuals. In the study from Taskar et al., correlations between individuals were constructed using user attributes with relational Markov Networks. For instance, they proposed a transitivity pattern that is useful in relationship prediction, where the presence of A-B relation

and B-C relation promotes the probability of A-C relation. Context information, such as locations and periods of time, has also shown potentials to predict social ties. Using location information alone may not be a sufficient predictor. In Crandall et al.'s work, only 0.1% of the relations were predicted with a confidence of 60%. However, when network structure are analyzed together with location information, over 90% friendship were detected with confidence over 80%, illustrated in the study from Sadilek et al.. Although location information alone is not a good indicator for friendship inference, a number of researches have indicated the importance between social ties and distance. The integration of location information and other features are also proved to be of high accuracy in friendship inference.

2.2 Community Detection

While friendship inferring are engaged to predict the local relationships between individuals, community detection, from a global aspect of view, groups people into smaller subgroups with tighter relationships. In previous research such as the studies from Mislove et al. and Xie and Szymanski, from people's friendship conditions in social network, community detection groups the people into overlap or distinct "communities", while the members of the same community known quite a few others from the same community or the members share the similar characteristics. Further more, studies from Mislove et al. showing that, community detection can be used to infer the vacant profile information of people based on the profile information from other members in the same community.

From the algorithm implementation view, the social network is usually defined as an undirected unweighted graph, while the vertexes represent the individuals, the edges represents the friendship between individuals. To change the social network into communities, various of problem statement and algorithm were posed. In the study of Girvan and Newman, the problem statement was " removing the edges which is likely to be the friendship across communities, until the left graph divided into unconnected components, while each component represent one community". On the other hand, in the study of Newman, the problem statement becomes " start with each individual as a community, then merge the community into another one with greatest increase on the global modularity".

Besides the classic community detection model, as our social network may contains friendship duplicate between individuals such as multiple phone calls between pair of individuals, we may also try to apply clustering algorithms on weighted graph representations.

3 Problem formulation

Describe your project as a machine learning problem, identify inputs objects, labels, possible features

3.1 Community Detection

For community detection problem, the ideal community result we want to get is the group belongings of the individual. As provided by the dataset we use, there were 94 volunteer partici-

pate the entire data collecting process. 68 of them were colleagues working in the same building on campus (90% graduate students, 10% staff) while the remaining 26 participants were incoming students at the universitys business school. The above identification information for each individual is stored inside the dataset. For the problem statement, we define the network between the participants as two distinct social network, phone call network and Bluetooth network. Each phone call or Bluetooth connection between individuals would be counted as an "edge" in the network. As both of the networks may contain duplicate friendships, we may firstly turns the relation condition between people as a binary variable, which is " whether the two individuals have connection". For the fundamental implementation, we would use the clustering and modularity algorithm. Then, for an advanced trial, we would combine the duplicate friendship as weighted edge, then apply K-means clustering on the entire network. Finally, we will try to combine the phone call network together with the Bluetooth network, and see if there is any improvement.

4 Data and Evaluation Plan

We will use a reality mining dataset from MIT media lab Eagle et al.. The dataset consists of phone logs of 94 subjects from September 2004 to June 2005. Among these 94 subjects, 68 were colleagues working in the same building (90% were graduate students, while 10% were staff). The remaining 26 subjects were incoming students from the business school. The dataset was collected from Nokia 6600 phones programmed to automatically run a log application as background process, including phone log, bluetooth and location. The format of each log is summarized as following:

* Phone log: (TIME) 20060720T211505 (DESCRIPTION) Voice call (DIRECTION) Outgoing (DURATION seconds) 23 (NUMBER) 6175559821

* Bluetooth: (TIME) 20060721T111222 devices: 000e6d2a3564 [Amys Phone]000e6d2b06ea [Jons PalmPilot]

* Location:(TIME) 20060721T111222 (CELL AREA) 24127, (CELL TOWER) 111, (SERVICE PROVIDER) AT&T Wirel (USER DEFINED LOCATION NAME) My Office

For the evaluation function of the community detection algorithm, we will check the group label inside each community, then compute the purity of the community, which is the purity of the cluster shown as $\frac{\#majority\ group\ member}{\#totalgroupmember}$. Then, we will make a weighted combination for the purity in each community according to the size of community, then use the final purity value as the evaluation of our community detection algorithm(A value between 0 and 1, the larger the better). For the baseline of the result, we should make the whole experiment group as one cluster, then compute the purity value for comparison.

How will you evaluate your algorithm? What is a reasonable baseline?

Submission Instructions:

delete this section when submitting

You are required to use LaTeX to type your solutions to questions, and report of your programing as well. Other formats of submission will **not** be accepted. A template named "homework.tex" is also provided for your convenience.

After logging into data.cs.purdue.edu (physically go to the lab or use ssh remotely, you are all granted the accounts to CS data machines during this class), please follow these steps to submit your assignment:

1. Make a directory named *'your Name_your Surname'* and copy all of your files there.
2. While in the upper level directory (if the files are in /homes/dan/dan_goldwasser, go to/homes/dan), execute the following command:

   `turnin -c cs578 -p PROPOSAL *your_folder_name*`

   (e.g. your prof would use: `turnin -c cs578 -p PROPOSAL dan_goldwasser` to submit his work)

   Keep in mind that old submissions are overwritten with new ones whenever you execute this command.

3. You can verify the contents of your submission by executing the following command:

   `turnin -v -c cs578 -p PROPOSAL`

   Do **not** forget the -v flag here, as otherwise your submission would be replaced with an empty one.

References

D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. 107:22436–22441. ISSN 0027-8424. doi: 10.1073/pnas.1006155107. URL http://adsabs.harvard.edu/abs/2010PNAS..10722436C.

Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. 106(36):15274–15278. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0900282106. URL http://www.pnas.org/content/106/36/15274.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. 99 (12):7821–7826. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.122653799. URL http://www.pnas.org/content/99/12/7821.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. 58(7): 1019–1031. ISSN 1532-2882. doi: 10.1002/asi.v58:7. URL http://dx.doi.org/10.1002/asi.v58:7.

Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. You are who you know: Inferring user profiles in online social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 251–260. ACM. ISBN 978-1-60558-889-6. doi: 10.1145/1718487.1718519. URL http://doi.acm.org/10.1145/1718487.1718519.

M. E. J. Newman. Fast algorithm for detecting community structure in networks. 69(6): 066133. doi: 10.1103/PhysRevE.69.066133. URL http://link.aps.org/doi/10.1103/PhysRevE.69.066133.

Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 723–732. ACM. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124380. URL http://doi.acm.org/10.1145/2124295.2124380.

Ben Taskar, Ming-fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In S. Thrun, L. K. Saul, and B. Schlkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 659–666. MIT Press. URL http://papers.nips.cc/paper/2465-link-prediction-in-relational-data.pdf.

Jierui Xie and B.K. Szymanski. Community detection using a neighborhood strength driven label propagation algorithm. In *2011 IEEE Network Science Workshop (NSW)*, pages 188–195. doi: 10.1109/NSW.2011.6004645.