Final Project Report

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

December 11, 2015

1 Introduction and Motivation

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 focused 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 3. Different methods are explained in section 4 and experimented (see section 5). Conclusions and discussions are presented in section 6.

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. show 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 also applied the transforming algorithm from the weighted and directed graph to the unweighted undirected graph for a better analysis result.

3 Data

A reality mining dataset from MIT media lab Eagle et al. was used in this project. 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

From this dataset, bluetooth and mobile information was utilized. As invalid mac address discovered in bluetooth data, 9 more subjects' data was eliminated. Therefore, a total of 85 subjects' data was used. In these 85 subjects, the friend relations are sparse. Observed from 1980 mobile connection data, only 31 entries of data is from individuals who are friend. Considering this dataset was originally aimed for community detection analysis, we treat the friendship inferrence as an assisting bridge to the community detection problem in this project.

4 Method Demonstration

4.1 Friendship Inference

To infer the friendship relations between a pair of individuals, according to different approaches, this problem can be formulated differently. In this project, we tried three different approaches, one method based on topology structure, one based on pairwise information and one integrating the former two.

4.1.1 Problem Formulation

- To utilize topology structure, this problem can be formulated as an input of network, representing the connections between pairs of individuals, with an output network indicating the friendship relation. The input network can be various, including the phone call connections, bluetooth networks and shared locations. In this project, the network of bluetooth scans is used as the input. Similarities, such as number of shared neighbors, between two vertices (individuals) in the network is computed to infer the friendship relation.

- Other than topology structure, we also tried machine learning algorithm, svm, to train a classifier using information between pair of individuals. Therefore, the input of this method is the sampled training data to train a classifier. With this classifier, we can get the output label of test data. This is a two-class classifier, label 1 indicating the pair of individuals are friends; otherwise not. Each training data contains the information between a pair of individuals, represented by a tuple of n elements, with first and second element to be the id of the two individuals, the last element to be the label indicating their relations and the rest elements to be the features. In this approach, mobile data was used and different combinations of features were tested, including the total number of phone calls, the total duration of phone calls and the ratio of night calls. Analysis and comparisons between features are discussed in section 5.

- A hybrid approach integrating both topology information and pairwise information is developed in this project. The input is a network of bluetooth scans, with the features extracted from mobile data. The number of shared neighbors from the input network is used as an additional feature and trained together with other features, using support vector machine in unbalanced case (explained in detail in section 5).

4.2 Community Detection

4.2.1 Problem Formulation

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 participate 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 a Bluetooth social network. Each Bluetooth connection between individuals would be counted as an "edge" in the network. As the network contains duplicate friendships, we firstly turned the duplicated relations between people into weighted and directed edges, which is " how many times did the individual A meet B and vice versa". After getting the weighted relations between people, we dropped the rare connections between individuals, which is the edges under weight of 5 in this case. Then, we changed the graph into an unweighted and directed graph.

From the problem definition of community detection, we know that the problem is about analyzing the topological structure of the graph. Thus, compared with sparse edges graph, we were more appealing to the graph with more edges, which means more information encoded potentially. As a result, when there was an asymmetric directed edge pair between two individuals, we claimed that they are related with each other, which means there was an edge between the two.

After all the pre-processing steps, we got an unweighted undirected Bluetooth graph as the input of Community Detection algorithm.

In this project, we defined the community detection problem as the problem defined in Girvan and Newman. Starting with the individual's Bluetooth relation graph, which is a single component graph, we continuously removed the edge which is the most likely to " across the communities" until the graph breaks into small multiple components. During the removing of the edges, whenever there was an increasing in the number of components, we would run the evaluation algorithm to check the purity and uniqueness of each component in current graph.

4.2.2 Algorithm Demonstration

To describe the likelihood for an edge to be the "cross community edge", we computed the "betweenness" of each edge inside the Bluetooth graph. For the computation of "betweenness", firstly, we compute the shortest path between each pair of vertices, then add 1 weight to each of the edges that used in this shortest path. When there are more than 1 distinct shortest path exist for a fixed pair of vertices, we added 1 weight for each of the edges that appear in at least one of the shortest path, to keep the global equality for the contribution of each shortest path.

After computing all the shortest paths inside the graph and adding the corresponding weight to the edges, we got the weight of each edge as the "betweenness". Then, for a single loop in the algorithm, we remove the edge with largest "betweenness", then check the number of the component, update the "betweenness" of the edges after edge removing, then start the next loop.

For a very simple version of implementation, as computing the shortest path from one vertex to all the other vertices needs $O(n^2)$ time, a one-time global computing of the "betweenness" of each edge requires $O(n^3)$ time. This means, to complete the whole process of edge removing, we need $O(n^5)$ computation time.

However, we kept a record for all the shortest path edges between each pair of vertices in this graph. As a result, after removing one edge, we would update only the shortest paths between the pair of vertices that had the removed edge included. This would drop the computation time significantly.

4.3 Relations between Frienship Inference and Community Detection

Friendship inference and community detection are two directions in the research of exploring social networks. These two directions can be different but also helpful in each other. For example, the level of friendship between individuals can also worked as a input network to predict community labels.

5 Result and Analysis

5.1 Friendship Inference

5.1.1 Feature Selection

To infer frienship relation, choosing the correct features can be an important aspect to prompt the classification performance. From mobile data, we extracted three features, including the total number of phone calls, the total duration of phone calls and the ratio of night calls. As a first thought, the ratio of night calls seems to be an attempting important indicator to friend relations. However, in experiment, none of the friend relations can be found if only this feature is used. Besides the ratio of night calls, the total duration of phone calls is also a good indicator for friends. Almost half of the friend relations can be detected, when using both durations and ratio of night calls. However, the feature of number of phone calls is not as useful as the other two features. When tested with all three feature used, there is no improvement of detecting friend relations, but only 3 more number of non-friend relations.

The importance of night call ratio feature can also be observed by eliminating this feature. When this feature is not used, only 12.9% of the friend relation was detected with an significant number of misclassified non-freind relations.

The non-friend relations is found to be easier to detect since the data contains much more samples of non-friend relations. 99.0% of the non-friend relations were successfully detected when three features are all used. The ratio of friend and non-friend relation is around 1 : 70. As the huge difference between the number of friend and non-friend data, an unbalanced version of svm is used to leverage this situation.

5.1.2 Comparison between three method

For the approach using topology structure, a precision of 57.90% with a recall of 7.19% was reached. The low rate of recall indicated the poor abilitiy of using only similarities between two nodes. For the approach using only pairwise information, a precision of 66.67% with recall of 32.26% was reached. And for the hybrid approach, a precision of 54.55% with recall of 38.71% was obtained. A higher accuracy was achieved when detecting friend relations using the hybrid approach. This advantage is more obvious when less feature was used. This implication can be furthur tested using a more proper dataset.

5.2 Community Detection

As demonstrated in the dataset instruction, all the participator of this experiment are label with one of following group name. The subjects affiliation is following:

- ’mlgrad’ Media Lab Graduate Student (not a first year)
- ’1styeargrad’ Media Lab First Year Graduate Student
- ’mlfrosh’ Media Lab First Year Undergraduate Student
- ’mlstaff’ Media Lab Staff
- ’mlurop’ Media Lab Undergraduate
- ’professor’ Media Lab Professor
- ’sloan’ Sloan Business School

From the name of the affiliation, we can observe that most of the group are belongs to the Media Lab, while the group of Sloan Business School has the most obvious differences with others. This means that, if the community detection algorithm work correctly, we should firstly break the whole component into two sub-community, while one of the community includes all the member of Media Lab School, the other includes all the membe of Sloan Business School.

For a more advanced expectation on the community detection result, the algorithm should break the Media Lab component into smaller sub-component, while each of the components includes all the members with the same affiliation.

As shown in the algorithm tracing result in 7, in the beginning of the edge removing and component breaking, the algorithm cut the small group from the major component, while each of the new sub-group remained the 100% purity inside itself. Then, when the original component broke into 5 communities, we saw a clear separation between the most of Media Lab member and all the Sloan Business School member. As the new components remained 100% purity, this could be valued as a high performance community detection algorithm if we treat all the member of Media Lab as one general community.

After separating all the member of Sloan School, the algorithm continued to cut small sub-group out from the original component, while all the new component remained 100% purity, but only had less than 3 members. This phenomena can be observed until we got the maximum number of components we set, which is 10 components. This means the Bluetooth behaviors across different sub-group inside Media Lab are almost identical with each other.

6 Summary

From the result of community detection, we can get the conclusion that the daily face-to-face meeting behaviors, which are represented by the Bluetooth scanning record, are similar among the members of the same school, while significantly differ between the members across the school. For general community detection, the edge removing method would have promising performance; but for the community detection among the groups which may closely related with each other, the algorithm would end up with cutting the large group into small pieces.

7 Appendix

The component breaking result is following:

* component: 1 edges removed: 0

  total: 85 max 30 { 'mlgrad': 30, 'professor': 1, '1styeargrad ': 15, 'sloan': 25, 'mlstaff': 4, 'mlurop': 2, 'grad': 2, 'mlfrosh': 6}

* component: 2 edges removed: 16

  total: 82 max 30 { 'mlgrad': 30, 'professor': 1, '1styeargrad': 15, 'sloan': 25, 'mlstaff': 4, 'mlurop': 2, 'grad': 2, 'mlfrosh': 3}

  total: 3 max 3 {'mlfrosh': 3}

* component: 3 edges removed: 6

  total: 81 max 29 {'mlgrad': 29, 'professor': 1, '1styeargrad': 15, 'sloan': 25, 'mlstaff': 4, 'mlurop': 2, 'grad': 2, 'mlfrosh': 3}

  total: 3 max 3 {'mlfrosh': 3}

  total: 1 max 1 {'mlgrad': 1}

- component: 4 edges removed: 44

  total: 80 max 29 {'mlgrad': 29, 'professor': 1, '1styeargrad': 15, 'sloan': 24, 'mlstaff': 4, 'mlurop': 2, 'grad': 2, 'mlfrosh': 3}

  total: 3 max 3 {'mlfrosh': 3}

  total: 1 max 1 {'mlgrad': 1}

  total: 1 max 1 {'sloan': 1}

- component: 5 edges removed: 68

  total: 56 max 29 { 'mlgrad': 29, 'professor': 1, '1styeargrad': 15, 'mlfrosh': 3, 'mlstaff': 4, 'mlurop': 2, 'grad': 2}

  total: 24 max 24 {'sloan': 24}

  total: 3 max 3 {'mlfrosh': 3}

  total: 1 max 1 {'mlgrad': 1}

  total: 1 max 1 {'sloan': 1}

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.