# CS 498: Project Midterm Report

Adit Krishnan, UIN: 659913017, Net ID: aditk2 , Aravind Sankar, UIN: 671514664, Net ID: asankar3

November 2, 2016

## 1 Introduction

The task of discovering social circles in a user's personal network is an important one, in the context of social media. The problem can be posed as a clustering task that aims to identify distinct groups of friends within the ego network of the target user. The goal of this project is to identify these circles in a target user's ego network provided the attributes of the ego user, and his friends, as well as the network structure of his friends (i.e edges between pairs of his friends). Ideally, we want to generate hard clusters that can overlap (since social circles do overlap in most real life contexts, a friend could also be a classmate). Soft clusters may not provide unambiguous assignments which is undesirable. Hence, we wish to develop a joint model using the network structure as well as the attribute information to predict overlapping social circles, and evaluate them against the ground truth circles which are available in the SNAP facebook dataset [3].

## 2 Related Work

Community detection is closely related to the problem of detecting social circles, since social circles can be conceived as distinct communities within a user's ego network. Classical node clustering algorithms attempt to cluster nodes to form communities either by considering graph structure in isolation or by matching attributes of users [9]. Models that are based on single assignment clustering [7, 8] fail to capture the inherent overlaps in circles. Alternately soft clustering methods attempt to probabilistically model the extent to which a node is part of a cluster. Topic models such as [5], [6] and variants thereof can be applied to identify soft clusters. [3] attempts to bridge this gap by modelling the network likelihood using both the attribute values and the community memberships (latent variable to be discovered).[4] identifies that the previous model may fail to capture the mutual dependency of the network structure of users. This is addressed by developing a common generative process to capure both.

## 3 Implementation

We implement 2 simple baseline algorithms as follows and present their evaluation results using the metric defined in [4] with $\delta$ set to Jaccard and F1-Score functions. While one of our baselines is purely based on the attribute values of the nodes, the other uses only the network structure. Thus ideally these baselines can be improved upon by a method that combines both these aspects of the network to find circles.

- **Baseline 1 - PLSA based community detection (Uses only attributes)**
  Topic Models can be trivially adapted to produce soft clusters among nodes in an attributed

network. Each node is treated as a document and it's attributes are treated similar to the words in the document. PLSA attempts to produce topics, which are esentially distributions over the attributes and cluster the nodes based on what their topic mixture values are. Once we obtain the soft clustering distributions of nodes, we can form hard clusters, by assigning each node to the cluster which has the highest component in the mixture. This method uses only attributes and not the network structure of nodes. Number of clusters was chosen to minimize perplexity.

- **Baseline 2 - Girvan Newman Clustering Algorithm (Uses only network structure)** The Girvan Newman clustering algorithm begins with all nodes lying in a single cluster. It then successively deletes the edges with the highest betweeness centrality. Eventually this leads to nodes being split into multiple components which form the clusters that are outputed. The dendrogram structure that is generated can be cut at different levels to obtain to different numbers of clusters of nodes.

## 4 Results

Both methods were evaluated against ground truth communities using the metric presented in [4] with Jaccard and F1 $\delta$ functions.

| | PLSA with $\delta$ = F1 | PLSA with $\delta$ = Jaccard | GNC with $\delta$=F1 | GNC with $\delta$=Jaccard |
|---|---|---|---|---|
| Ego Network 1 | 0.0830323714608 | 0.0861349592883 | 0.185326526504 | 0.166765973108 |
| Ego Network 2 | 0.0482687618024 | 0.0661493849308 | 0.249450226708 | 0.260856069586 |
| Ego Network 3 | 0.055392811404 | 0.0640557924006 | 0.269306877669 | 0.27004376444 |
| Ego Network 4 | 0.0556918084486 | 0.0552375845927 | 0.286015290551 | 0.139101789392 |
| Ego Network 5 | 0.0559724019111 | 0.0595146353208 | 0.0866878857066 | 0.084590761101 |
| Ego Network 6 | 0.089147290346 | 0.118093111177 | 0.210690622785 | 0.242270747497 |
| Ego Network 7 | 0.280912655971 | 0.250993973347 | 0.345375067924 | 0.244668691555 |
| Ego Network 8 | 0.0996365890523 | 0.134004097054 | 0.526744499216 | 0.38760800194 |
| Ego Network 9 | 0.109448298689 | 0.138776451434 | 0.286715595986 | 0.275998119974 |
| Ego Network 10 | 0.21588237209 | 0.20895802915 | 0.467940047814 | 0.253812985106 |
| Average | 0.109338536118 | 0.11819180187 | 0.291425264086 | 0.23257169037 |

## 5 Interpretation

- The PLSA based method produces poorer results. This indicates that it is hard to produce good communities in the absence of the links between nodes. Attribute information alone does not appear to generate good communities.

- A possible reason for the poor performance of the above baseline is sparsity in attributes. In most ego networks, a very large number of attributes are 0 for most users. A simple way to deal with this is to perform feature selection or feature extraction.

- Girvan Newman clustering is intractable for larger ego networks. In our facebook dataset [3], the larger ego networks take about 2 hours to cluster with an efficient Girvan Newman implementation from [1].

# References

[1] Leskovec, Jure. ”Stanford network analysis package (snap).” URL http://snap. stanford. edu (2013).

[2] Leskovec, Jure, and Andrej Krevl. ”SNAP Datasets: Stanford large network dataset collection, June 2014.” URL: http://snap. stanford. edu/data (2014).

[3] McAuley, Julian J., and Jure Leskovec. ”Learning to Discover Social Circles in Ego Networks.” NIPS. Vol. 2012. 2012.

[4] Yang, Jaewon, Julian McAuley, and Jure Leskovec. ”Community detection in networks with node attributes.” 2013 IEEE 13th International Conference on Data Mining. IEEE, 2013.

[5] Hofmann, Thomas. ”Probabilistic latent semantic indexing.” Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.

[6] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” Journal of machine Learning research 3.Jan (2003): 993-1022.

[7] L. Akoglu, H. Tong, B. Meeder, and C. Faloutsos. PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs. SDM 12, 2012.

[8] M. Ester, R. Ge, B. Gao, Z. Hu, and B. Ben-Moshe. Joint Cluster Analysis of Attribute Data and Relationship Data: the Connected kCenter Problem. In SDM 06, 2006.

[9] Johnson, Stephen C. ”Hierarchical clustering schemes.” Psychometrika 32.3 (1967): 241-254.