Social Circle Analysis

Vignesh Kannan
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
vkannan2@andrew.cmu.edu

Siddharth Kannan
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
siddhark@andrew.cmu.edu

1 Motivation

Social networking sites have transformed information sharing and connectivity between people in the last decade. With the emergence of various platforms like Facebook, Twitter and Google+ there has been an explosion of data associated with these social networks and an increasing interest to model them. It is natural to think of social networks as being composed of many interconnected social circles and indeed, this has been an active area of research with numerous applications like filtered sharing of content, marketing, targeted advertisements and even for the purposes of international security. Additionally, we can also infer missing information about users from their memberships to different social circles and improve friend and page suggestions to them. This can also facilitate the formation of online communities and forums to mobilize support for various causes. Finally, the study of the formation of these social circles can be of significant importance from the standpoint of social sciences too as we can gather new insights about the interconnected and interwoven nature of society itself.

2 Related Work

The most common and obvious way to model a social network is in the form of a graph, where the nodes are the users and the edges are the connections between the users. From our study, we could broadly classify the approaches used into two categories. The first approach deals with the modeling of what is called as the ego network, wherein the goal is to construct social circles around a specific user referred to as the ego. Leskovec et al. (1) model the probability of existence of an edge between two vertices. They treat the membership of vertices to the circles as latent variables and propose an Expectation Maximization (EM) like framework to alternately optimize circle membership and the parameters of the user profile similarity function until convergence. This enables the incorporation of both structural and node content information in the model and the parameters learned are also interpretable in the sense that one can discern the profile attribute along which a circle has emerged.

The second approach operates from a global perspective wherein we are interested in modeling social circles on the whole graph. To this end, Wang et al. (2) propose an approach that uses a stacked graph autoencoder to combine node content and graph structure in the spectral domain and performs a spectral clustering on the learned representation. Additionally, they add random noise to the node content in order to capture the dynamic interplay between node content and structure and are also able to achieve a closed form solution for the autoencoder.

Another approach is proposed by Yang et al. (3) wherein the main focus is to address the problem of missing information which could either be missing links between two users or missing attributes of a user. The solution here is based on the *homophily* phenomenon which states that users with similar attributes tend to form links between each other and vice versa. In a unified probabilistic framework, the authors propose an iterative algorithm composed of two alternating steps: Graph Construction (GC) and Label Propagation (LP). GC tries to estimate the probability of link formation

Preprint. Work in progress.

between pairs of users given some already existing links and the attribute information of users. LP tries to infer attributes for users based on some already given attributes and probability of link formation (affinity) between pairs of users. The two steps alternate with each other until convergence. The ability to complete missing information, apart from being useful in finding clusters, has many potential applications like detecting anomalies in link structure by comparing the completed links to the existing network, and content targeting based on inferred attributes to improve business value.

3 Methodology

3.1 Desirable Properties

We have identified the following properties that we would like our model to have in order to capture the true nature of the organization of social circles and their connectivity.

1. Social circles evolve in a hierarchical bottom-up fashion.
2. Circles can overlap or in other words, a user can belong to more than one circle.
3. Circles can form within circles.
4. Principle of homophily is captured by the framework.
5. Social circles are dynamic in nature as they constantly get modified by the addition of new friends to a user or change in attributes of the user.
6. Some attributes result in the formation of denser circles than other attributes.
7. Users with many mutual friends are likely to be connected.

3.2 Approach

We would like to consider a non-parametric framework for modelling this problem since we believe that social networks can have arbitrarily complicated structures and using parametric assumptions would affect the span of our model. Motivated by the randomness in how real world social circles have formed over time, we propose a randomized algorithm to identify social circles given an input social network graph with attribute information. We start off in a hierarchical fashion by forming small clusters between nodes with highly similar attributes. We then allow the clusters to grow and more importantly overlap (which is not accounted for by simple hierarchical clustering) in a randomized manner. More precisely, we associate a probability value with the overlap of two clusters. Based on this probability we choose to create appropriate overlaps between pairs of clusters which have at least one edge between them.

Features which we would like to account for while computing the overlap probabilities are the similarity between cluster attributes and the link structure between the two clusters. Once we have consumed all the edges in the graph, we perform a Label Propagation step as done by Yang et al. (3). The rationale behind this step is two-fold. Firstly, this lets us infer missing information for user nodes. Secondly and more importantly, it lets us update the overlap probabilities between two existing clusters. We then use these updated probabilities to again randomly simulate the overlap of two clusters as done previously. These two steps (Random Overlap and Label Propagation) alternate till the overlap probabilities converge to values close to either 0 or 1. We can use functions like the sigmoid (since the tails of the sigmoid asymptotically approach 0 or 1 quite rapidly) to accelerate this convergence. If time permits, we could also try to achieve formal guarantees on the convergence.

As an aside, we consider the usage of the Disjoint-Set (Union-Find) data structure to implement the formation of clusters and simulate their overlap. A challenge in this step is to somehow incorporate the notion of overlap in the data structure which is inherently designed to work with disjoint sets.

References

[1] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, pp. 539–547, 2012.

[2] C. Wang, S. Pan, G. Long, X. Zhu, and J. Jiang, “Mgae: Marginalized graph autoencoder for graph clustering,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 889–898, ACM, 2017.

[3] C. Yang, L. Zhong, L.-J. Li, and L. Jie, “Bi-directional joint inference for user links and attributes on large social graphs,” in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 564–573, International World Wide Web Conferences Steering Committee, 2017.