2012 Link Analytics Challenge: The Invisible Edge

November 14, 2012

1 Introduction

It is simple to model social networks as graphs: the nodes are people, edges indicate communication between two people. In some settings, we are able to observe the entire social network. For example, each of you know who your friends are, you know who your Facebook friends are, you know who you communicate with via text or email, you know who you communicate with via voice communication. However, as networks become larger, it becomes more and more difficult to observe the entire graph and there may be nodes and edges that are either difficult or impossible to observe.

In this project, you will be given two sets of graphs. The first set of graphs is a true representation of the network. However, in the second set of graphs, some edges will have been removed by us, making them effectively invisible. It is your task to tell us whether or not an edge between two nodes is invisible!

2 The Networks

At the beginning of the project you will be given two large graphs that are representative of the communication network of two large, yet very different, U.S. cities. Each graph represents one month of mobile communication where at least one party belongs to the city in question. In these graphs, there are no missing edges. The nodes will be labeled with a unique identifier, and the edges will be labeled to represent various aspects of the relationship.

The main focus when studying these networks is to become familiar with various structural properties of these networks, and to develop some kind of predictive capability: When I see a pair of nodes that has some property x, there is an edge between these nodes with probability p_x. It is your task to decide what kind of properties are relevant and to try to estimate these probabilities. Here are some possible things to think about:

- There are many triangles in the graph. Intuitively, your friends are friends of one another. If you only observe two sides of a triangle, can you predict the presence of the third edge based on the properties of the other two? Can other community structures improve your ability to predict the presence of such edges?

- The degree distribution of these types of graphs are well-studied: while most people have only a handful of friends, there are others with hundreds, even thousands of "friends". How can this information be used?

- Large cliques exist in the graph - what can you say about these cliques? Are the relationships in these cliques unusually strong or intense?

- Is the graph connected? If not, what can be said about the different components? Does this tell us anything about communities in the network?

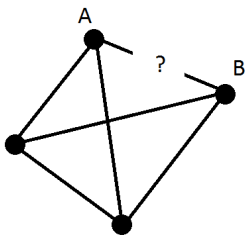

Figure 1: A and B have friends in common who know each other. Are A and B friends?

- Visualizing large graphs is notoriously difficult - can you come up with any visualization techniques that help you understand these networks to help your ability to predict invisible edges?

In the second part of the project, you will study the same social networks for these two cities for the next month. However, the nodes will have different labels so that a node with the label 1234 in month one will not be labeled 1234 in the month two network. With some effort, you could probably figure out how the nodes were re-labeled for month two, but we are trusting that you will not waste your time on this counter-productive activity! For these month two networks, a large set of edges will be made *invisible* according to a random process. We will provide you with a large list of candidate edges that may or may not have been removed. Your task will be to differentiate between legitimate edges that were made invisible and random edges (more on this later).

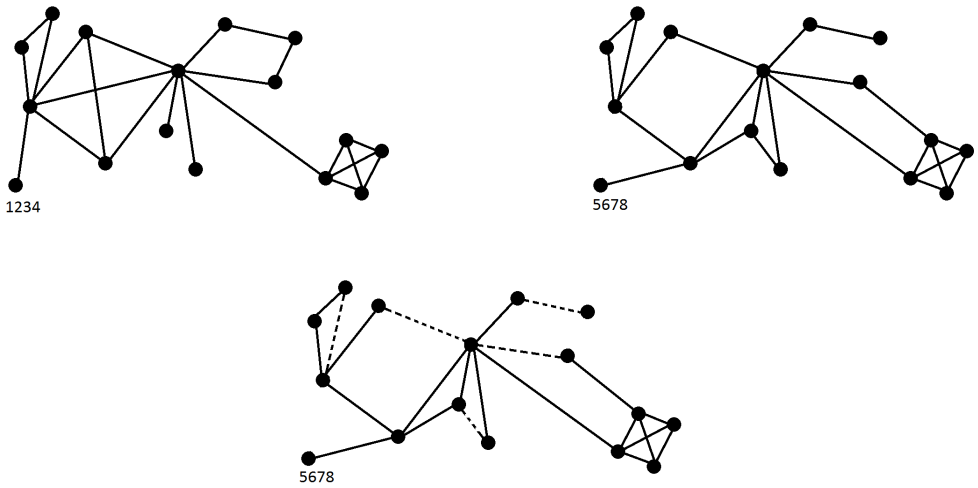

Figure 2: Example network for month one, month two, and then month two after we remove some edges (removed edges are dashed). While the bottom left node is labeled 1234 in month one, it has a different label, 5678, in the second month.

3 The Challenge

The networks are given as a labeled edge list. Each line in the file corresponds to a relationship between two entities, A and B, formatted as a space-delimited list as follows:

A	A type	B	B type	# days	# calls	# secs	# texts
188019861655694	1	69428241743694	0	3	1	12	3

Table 1: Example line from network file

The A and B fields are simply unique numeric identifiers assigned to the nodes in the graph. A type and B type are either 0 or 1. We will reveal the meaning of these fields at the end of the challenge. However, any insight you can extract in terms of differences in behavior between type 0 and type 1 nodes will be very interesting! # days indicates how many distinct days A and B communicated during the month. The remaining fields should be self-explanatory. We will provide a README file to accompany the data to ensure that there is no confusion.

In order to judge your skill in understanding these social networks, your will be asked to identify likely edges and tell us when two entities "should" know each other. We will provide you with a large list of edges that may or may not have been removed from the networks of the second month of data. You will be provided with a simple, two column file where each line in the file indicates an edge between two nodes in the graph. For each edge (A, B) in this file, there are two possibilities: 1) either the edge existed in the original network existed and we made it invisible, or 2) the edge did not exist in the original network. For each edge, your task is to tell us whether or not you expect that edge to exist, by assigning a 0 or 1 to the edge. You will be judged on the accuracy of your predictions. The final output for this should be a space-delimited 3-column file formatted exactly like this:

- A B 0/1

Again, a 1 indicates that you think A and B really do know each other and that this edge was made invisible, and a 0 indicates that you believe this is a randomly generated edge and you have no strong evidence to indicate that A and B are "friends." We can tell you that roughly half the edges are random, and the other half are legitimate edges that were made invisible.

The overall challenge winner will be determined by assessing your accuracy in predicting when edges "should" exist. However, bonus points will be given to groups that uncover particularly novel or interesting properties of these graphs. Any insight you have into different behavior of the type 0 nodes versus the type 1 nodes will be looked upon very favorably. Also, any observations on how the networks for these two cities are different will also be very relevant.

Good luck!