CSC 591 – GDM Fall 2016

Homework 6 – Christos Faloutsos Lecture Assignment

Unity ID: avshirod

Questions: (Based on this lecture)

1. What evidences are provided in support of the claim that real world graphs are NOT random?

If a large enough data set is given, we can discover some sort of pattern in it. (Whether it can be reasoned or not is a different question.)

But phenomenon like 6-degrees of separation (Bacon rule, Milgram experiment) demonstrate that large real world graphs do turn up with certain similar features; and hence are not random.

One of the examples given was the Yahoo WebGraph (1.4 Billion nodes, 6.6 Billion edges), which in turn showed an effective diameter of about 7 in the undirected graph.

2. What distribution does the number of triangles in social networks resemble? Is there any correlation between the node degree and the number of triangles the node belongs to?

Number of triangles in social network follows a line, with skewed tail at the end (on a log-log scale).

Number of triangles and the degree of a node have a strong linear correlation. (More neighbors will lead to more triangles).

3. How to quickly estimate the number of triangles in the graph?

$$Number\ of\ Triangles = \frac{1}{6} * Sum({lambda_i}^3)$$

We only need to consider top 10 or so eigenvalues for this, because the eigenvalue distribution gets skewed after that, and as we are cubing them, the difference would be visible only at very high precision.

4. What interesting phenomenon was found using EigenSpokes, what is its possible explanation, and what graph mining task is it related to that we covered in the course?

Plotting the mobile social graph (caller-callee data) by projecting it onto first two eigenvectors of its eigen-decomposition, gives data points that run along the axes. (We would've expected a nice spread with some exceptions towards the end). This was signified as '*EigenSpokes'*.

After plotting the tail users (who called a lot of times), and observing the adjacency matrix, it was spotted that this group of users called each other very much. They had the same neighborhood. The adjacency matrix was dense, which showed high connectivity. So using something like this, we can identify good communities in large graphs.

We observed something similar when we studied Singular Value Decomposition. We observed that the first two eigenvectors covered majority of the spread. By observing patterns such as EigenSpokes, we can easily learn many things about the data that are unknown otherwise.

5. What evolution of the graph diameter has been identified for time evolving graphs and what evidence is provided in support of that claim?

The diameter of a time evolving graph shrinks as the size of the data grows. As more and more connections will be formed over time, the diameter will shrink as the distant nodes will come closer through newly formed connections.

The example given in this case used various snapshots over years of the patent citation network. A new patent has to cite all patents which correspond to that particular topic. As time passes, the number of citations grows and the networks comes closer and closer. In the graph shown in the lecture, the diameter of the aforementioned network shrunk from 35 in 1975, to almost 10 in 2000.

6. How the largest disconnected community change in time evolving graphs? Do they shrink, grow, or stabilize?

The largest disconnected community (i.e. the second largest connected component) shows all above characteristics –

It shrinks because it gets absorbed in the largest connected component.

It grows over time, before getting absorbed.

It stabilizes (oscillates) within certain limits during its lifetime.

7. Does popularity of the blogs drop off exponentially? If not, then how?

It is right to assume that it would decrease exponentially, because the post would be most popular after its creation; and then fade away as it gets old. But it was observed that it rather fits the power law when plotted on a log-log scale.

8. What can be said about the duration of the phone calls? What is TLaC distribution?

The duration of phone calls follows a weird curve – Lower number of phone calls for low durations, high number of phone calls for medium durations, and again low number of calls for longer durations. But, the start of the distribution is skewed. There is a very high observance for calls with 0-1 second duration.

This parabolic shape is best explained by the log-logistic distribution of TLaC.

TLaC is The Lazy Contractor distribution. It is named because the concept is analogous to a lazy contractor. 'If some task has taken a long time, it will take even longer in future.' So if a contractor does not finish the house on time, it will most likely take a lot longer to actually finish.

9. What is OddBall algorithm good for? What information does it use?

OddBall algorithm is good for spotting anomalies in weighted graphs.

In the example given in the video, OddBall algorithm is used to observe patterns in PostNet, which a graph of blog posts.

OddBall algorithm uses the ego-net of a node and certain corresponding features, such as – degree of node, number of edges in ego-net of node, total weight of ego-net of node, principal eigenvalue of weighted adjacency matrix for ego-net of node. Then it compares it with other nodes in the graph.

10. How could fraud detection on eBay be captured according to the lecture?

A fraud person creates multiple fake accounts, and gives positive feedback to his own fake accounts, thus falsly identifying himself as a trustworthy seller. How to detect this? This closely connected network between false accounts is similar to a nearly complete bi-partite clique.

11. How are the following questions answered: (a) Which nodes to immunize? (b) will a virus vanish or will it create an epidemic?

(a) We immunize the nodes which will maximally raise the epidemic threshold.

(b) Epidemic threshold depends on network connectivity. We can determine the epidemic threshold based on the reciprocal of the first eigenvalue in the network's adjacency matrix. If the strength of the virus (= Attack Prob (Beta) / Healing Prob (Delta)) is above threshold, the virus will turn into an epidemic, otherwise it will die quickly.