SS-ZG548: ADVANCED DATA MINING

16 Mining with Social Data

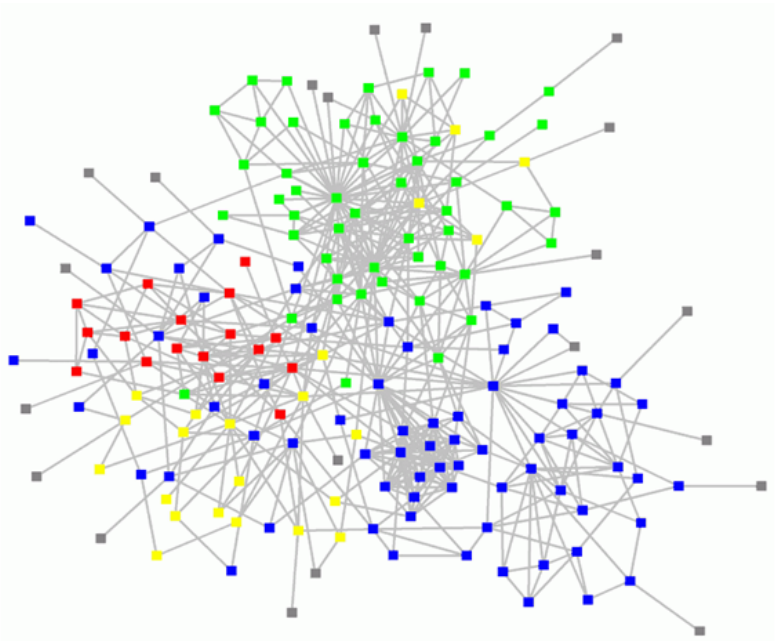
Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Oct 29, 2021 ONLINE (WILP @ BITS-Pilani July-Dec 2021)

http://ktiwari.in/adm

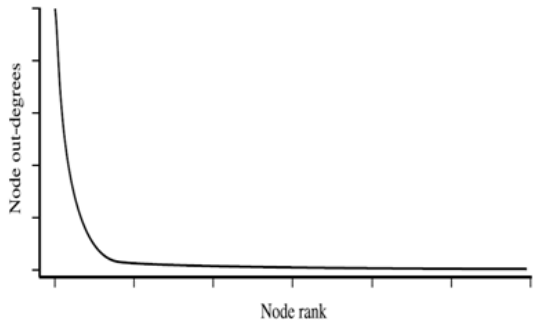
Social Networks

- Social network analysis (SNA) is the study of social networks to understand their structure and behavior
- Social Networks referred to technically as a graph. Each person is represented as a node. Connections are called links or edges
- What qualities can we look at? 1) Nodes degrees 2) Number of edges incident on a node 3) network diameter 4) maximum distance between pairs of nodes 5) Average distance between a pair of nodes 6) Shortest path length 7) Effective diameter



Characteristics of Social Networks

- Social Networks are rarely static
- Densification power law: Believed that as a network evolves, the number of degree grows linearly in the number of nodes. But it is found to be super linear in the number of nodes $e(t) \propto n(t)^a$
- Shrinking Diameter: Effective diameter tends to decrease as the network grows.
- Heavy-tailed out-degree and in-degree distributions: The number of out-degrees for a node tends to follow $1/n^a$, where n is the rank of node in the order of decreasing out-degrees ($0 < a < 2$). New node attaches by a constant number of out-links.



Link Mining

- **Link Prediction problem:**[1] Given a snapshot of a social network at time t, we wish to predict the edges that will be added to the network during the interval from time t to a given future time t'
- A social network $G = (V, E)$, each edge $e(u, v) \in E$ represents an interaction between u and v at time t(e)
- Multiple interactions between u and v as parallel edges at different time stamps
- For two times $t < t'$, let $G[t, t']$ denote the subgraph of G consisting of all edges with a time-stamp between t and t'
- Choose four times $t_0 < t_0' < t_1 < t_1'$, and give an algorithm access to the network $G[t_0, t_0']$; it must then output a list of edges, not present in $G[t_0, t_0']$, that are predicted to appear in the network $G[t_1, t_1']$. The interval $[t_0, t_0']$ refered as training interval and $[t_1, t_1']$ as testing interval

[1] Liben-Nowell, David and Kleinberg, Jon, "The link-prediction problem for social networks" In Journal of the American society for information science and technology, volume=58(7), pages=1019–1031, Wiley Online Library 2007

Link Mining

- Social networks grow through the addition of nodes as well as edges not sensible to seek predictions for edges whose endpoints are not present in the training interval
- Two parameters $K_{training}$ and K_{test} are used
- Define the set Core to be all nodes incident to at least $K_{training}$ edges in $G[t_0, t_0']$ and at least K_{test} edges in $G[t_1, t_1']$
- Evaluate how accurately the new edges between elements of Core can be predicted
- Each link predictor p that we consider outputs a ranked list L_p of pairs in $A \times A$; these are predicted new collaborations, in decreasing order of confidence
- **For evaluation**
  - Consider Set core, so we define $E^*_{new} = E_{new} \cap (core \times core)$ and $n = |E^*_{new}|$
  - Performance measure for predictor p is as follows: from the ranked list L_p, we take the first n pairs in $Core \times Core$, and determine the size of the intersection of this set of pairs with the set $|E^*_{new}|$

Link Prediction-Methods

- Based on Node Neighborhood: two nodes x and y are more likely to form a link in the future if their sets of neighbors $\Gamma(x)$ and $\Gamma(y)$ have large overlap

$$score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

- Verifying a correlation between the number of common neighbors of x and y at time t, and find the probability that they will collaborate in the future
- Measures the probability that both x and y have a feature f, for a randomly selected feature f that either x or y has

$$score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

- Other methods
  - Based on the ensemble of all paths
  - Random Walk
  - SimRank

Thank You!

Thank you very much for your attention!

Queries ?