COMS E6998-2: Analysis of Social Information Networks — Apr 7, 2011

Lecture 10 – Paper Presentations

Instructor: *Augustin Chaintreau* — Scribes: *Yoonji Shin*

1 Measurement and Evolution of Online Social Networks

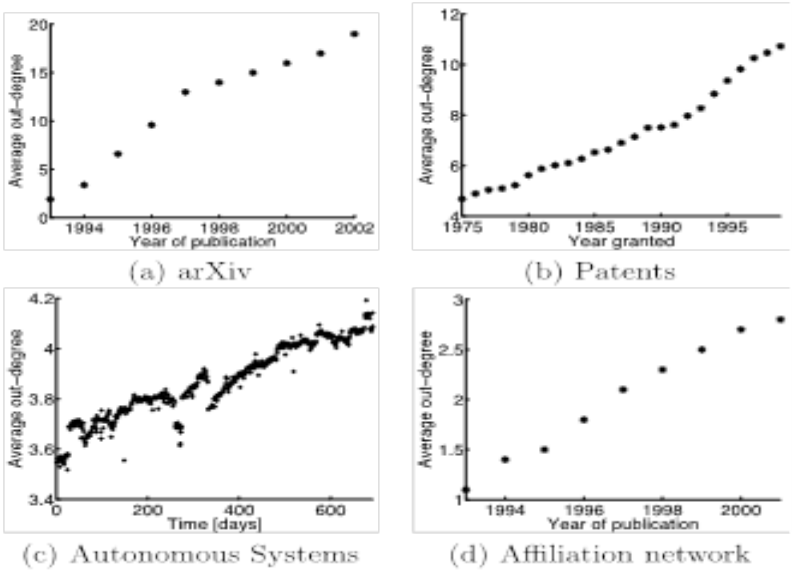
1.1 Motivation

This paper focuses on observing the evolution over time and defining the normal gorth patterns.

Recently, lots of studies have focused on understanding the social, technological, and scientific networks, and observed that the densification over time (super -linearity of edges) follows the quation $e(t) \backsim n(t)^a; a \in (1,2)$, and that the average distance shrinks over time. Studying the densification laws, shrinking diameters and possible explanations by observing graphs over time will allow us to generate graphs for simulations, and help alleviating bad sampling in graph sampling, extrapolate between past and future, and detect abnormality in more effecient and accurate manner.

1.2 Observation 1: Densification Law Exponents

The following graphs are based on 9 datasets from 4 different sources.



(a) arXiv (b) Patents

(c) Autonomous Systems (d) Affiliation network

Example 1. *a. Arxiv Citation Graph*

The data was collected between January 1993 to April 2003 by taking monthly snapshots, and Directed (i, j) in E means paper i cites paper j. Densification Law plots exponents $\alpha = 1.68(\gg 1)$, deviation from linear growth. The phantom nodes and edges phenominon is due to mising past.

Example 2. *b. Patents Citation Graph*

Data is based on published papers from January 1, 1963 to December 30, 1999. For densification law plotting the exponent is $\alpha = 1.66$

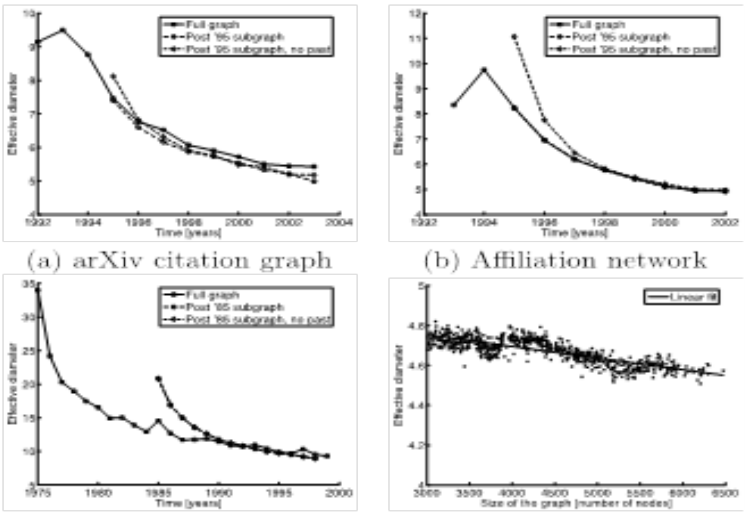
Example 3. *c. Autonomous Systems Graph*

This graph shows communication network between Autonomous Systems(AS) and te traffic between peers. The data was collected between November 8, 1997 to January 2, 2000. Densification Law plots exponents $\alpha = 1.18$

Example 4. *d. Affliation Graph*

This graph a bitartie graphs from ArXiv. Edges represent the connection between people to the papers authord, with co-authorship. The data was collected between April 1992 to March 2002. $\alpha \backsim 1.08 - 1.15$

1.3 Observation 2: Shrinking Effective Diameter



(a) arXiv citation graph (b) Affiliation network

(c) Patents (d) AS

In graph (a), we can observe the phantom nodes and edges, which are due to missing past. We can pick $t_0 > 0$ as the beginning of data and put back in the nodes and edges from before time t_0

The solid line represents full graph, which is effective d of the graphs' giant component. Dotted line with black nodes represent the post '95 subgraph as if we knew full past, and the dotted line with empty nodes represent post '95 subgraph with no past, which deletes all out links.

We can also observe that the diameter of GCC shrinks in ER model. However, this does not lead to shrinking diameters.

1.4 Proposed Models

The paper proposes models to capture the phenomina of densification power law, shrinking diameters, heavy tailed in-out degree distributions, and small world phenomena. $e(t) \backsim n(t)^a$ because each new node $\backsim n(t)^a$ out-links no insight.

Examples of recursive patters include geography based links in computer networks, smaller communities easier ties in social nets, and hyperlinks for related topics WWW.

1.4.1 Community Guided Attachment (CGA)

For a tree, let H be height and b be branching factor. Then number of leaves $n = |V|; n = b^H$ h(v, w) represents the height of smallest sub-tree containing both v, w.

We can see that the difficulty function (linkage probability) decreases in h, and is scale free. Thus,

$$\frac{f(h)}{f(h-1)} = const \rightarrow f(h) = f(0)c^{-h}; f(0) = 1, c \geq 1 \tag{1}$$

Theorem 5.

In the CGA random graph model, the expected average out-degree d of a node is proportional to
$d = n^{1-log_b(c)}$ if $1 \leq c < b$
$d = log_b(n)$ if $c = b$
$d = constant$ if $c > b$

1.4.2 Dynamic Community Guided Attachment

We can construct this by, in time step t, adding b new leaves as children of each leaf: b-ary tree of depth (t 1), and form out-links with any node independently. The tree-distance d(v,w) is the length of path between v,w in tree. e.g. for leaves: d(v,w) = 2h(v,w)

Theorem 6.

The dynamic CGA model has the following properties:

1. When $c < b$, the average node degree is $n(1 - log_b(c))$ in-degrees $\backsim$ Zipf distribution with exponent $\frac{1}{2}log_b(c)$

2. When $b < c < b^2$, the average node degree is constant, and the in-degrees $\backsim$ Zipf distribution with exponent $1 - \frac{1}{2}log_b(c)$

3. When $c > b^2$, the average node degree is constant and the probability of an in-degree exceeding any constant bound k decreases exponentially in k

1.4.3 Forest Fire Model

The idea is that we link v_i where i = 1, ..., k to odes of G s.t. G_{final}. Then we can observe "rich get richer" attachment result in heavy-tailed in-degrees, copying model results in communities, community guided attachment results in densification power law, and shirinking effective diameters.

We can construct this model by

i. nodes arrive in over time

ii. center of gravity in some part of the network

iii. linkage probability decreases rapidly with their distance from this center of gravity

The result of this model is very large number of out links skewed out-degree distributions and "bridges" for disparate sub-graphs.

1.5 Conclusions

By studying time evolution of real graphs, we found that

1. Densification Power Law averages out-degree grow

2. Shrinking diameters in real networks, contrary to standard assumption, may exhibit a gradual decrease as the network grows

3. Community Guided Attachment model can lead to the Densification Power Law, only one parameter suffices

4. Forest Fire Model, based on only two parameters, captures patterns observed both in previous work and in the current study: heavy-tailed in- and out-degrees, the Densification Power Law, and a shrinking diameter

This study can potentially be applied to what if scenarios, forecasting of future parameters of computer and social networks, anomaly detection on monitored graphs, designing graph sampling algorithms, and creating realistic graph generators

2 Discussion in Measurement and Evolution of Online Social Networks

There are different forms, contexts, spheres of interaction on different platforms abd networks. Each network has different rules of interaction and different risks and rewards for linking or disconnecting in any given network. When observing daily reach of different webpages, we can observe different patterns in different social networks. For instance, google shows high traffic during workdays while facebook shows high traffic during the weekends. Also, YouTube shows very high traffic during holidays. These differences reflect the type of interaction and user behavior in each social networks.

Normal growth patterns in social, technological, and information networks show:

1. Constant average degree assumption: The average node degree in the network remains constant over time. (Or equivalently, the number of edges grows linearly in the number of nodes.) This can be explained by "Densification power laws" – The networks are becoming denser over time, with the average degree increasing (and hence with the number of edges growing super-linearly in the number of nodes). Moreover, the densification follows a power-law pattern.

2. Slowly growing diameter assumption: The diameter is a slowly growing function of the network size, as in small world graphs. This can be explained by "Shrinking diameters" – The effective diameter is, in many cases, actually decreasing as the network grows.

From map of sceince, we can observe that seemingly distant and unrelated scientific fields are in fact, all connected.

Observing paper citations and patents allow us to understand how each papers and patents are connected to other works, but there are implications of citing more and more papers and patents as every paper can only contain a finite number of references (possible bounded problem), and as there are more references to other patents, it become longer with longer prosecution time and longer office actions

3 Affiliation and Groups in Social Network

3.1 Motivation

The motivation is to provide a simple, realistic and Mathematically tractable model which are algorithmically useful that can explain all properties of social networks, densification, and shrinking diameter.

By observing the internet and web graphs, we can find that

1. The degree distribution of the Internet graph is heavy-tailed, and roughly obeys a power law.

2. A network evolves by new nodes attaching themselves to existing nodes with probability proportional to the degrees of those nodes.

3. Besides power-law degree distribution, the web graph consists of numerous dense bipartite subgraphs.

Thus, preferential attachment and edge copying are two basic paradigms that can both lead to heavy-tailed degree distributions and small diameter.

There already exists a model of small world graphs –Kleinberg–, which observes transitive friendship and is a nice starting point to analyze social networks, but it is not applicable in developing an understanding of real social networks, and is a static model. Also, Leskovec's model shows densification and shrinking diameter but is a complex model and does not powerfully analyze degree distribution, densification, and shrinking diameter simultaneously. Thus, this study is dedicated in creating a model that can better represent the affliation networks.

3.2 Proposed Model: Affiliation Networks

Actors and societies are two types of entities in social networks that are related by affiliation of the actors in the societies. Affiliation networks are the social network among the actors that results from the bipartite graph. Affliation networks can be obtained by replacing paths of length two in the bipartite graph among actors by an (undirected) edge. This process is called "folding" the graph. When the affliatoin network B and its folding G on n vertices is produced, B has a power-law distribution, and G has a heavy- tailed degree distribution, and the effective diameter of G stailizes to a constant value.

Using this model, in large random set R, we can sparsify the graph to have a linear number of edges while preserving all shortest distances to vertices in R and stretching distances by no more than a factor given by the ratio of the expected degree of actor and society nodes in the affiliation network.

3.3 Constructing Affliation Networks Model

3.3.1 B(Q, E)

B(Q, E) can be constructed by the following:

1. Fix two integers cq, $cu > 0$, and let $\beta \in (0, 1)$.

2. At time 0, the bipartite graph $B_0(Q, U)$ is a simple graph with at least cqcu edges, where each node in Q has at least cq edges and each node in U has at least cu edges.

3. At time $t > 0$, With probability β, a new node q is added to Q.

4. (Preferentially chosen Prototype) A node $q \in Q$ is chosen as prototype for the new node, with probability proportional to its degree

5. (Edge copying) cq edges are copied from q; that is, cq neighbors of q, denoted by u1, . . . , ucq , are chosen uniformly at random (without replacement), and the edges (q, u1), , (q, ucq) are added to the graph.

6. (Evolution of U) With probability $1 - \beta$, a new node u is added to U following a symmetrical process, adding cu edges to u.

3.3.2 G(Q, E)

G(Q, E) can be constructed by the following:

1. Fix two integers cq, cu, $s > 0$, and let $\beta \in (0, 1)$.

2. At time 0, the bipartite graph $G_0(Q, E)$ consists of the subset Q of the vertices of $B_0(Q, U)$, and two vertices have an edge between them for every neighbor in U that they have in common in $B_0(Q, U)$.

3. At time $t > 0$, With probability β, a new node q is added to Q.

4. (Edges via Prototype) An edge between q and another node in Q is added for every neighbor that they have in common in B(Q,U) (note that this is done after the edges for q are determined in B).

5. (Edges via evolution of U) With probability $1 - \beta$, a new edge is added between two nodes q1 and q2 if the new node added to $u \in U$ is a neighbor of both q1 and q2 in B(Q,U).

6. (Preferentially Chosen Edges) A set of s nodes q_i1 , . . . , q_is is chosen, each node independently of the others (with replacement), by choosing vertices with probability proportional to their degrees, and the edges (q, q_i1), . . . , (q, q_is) are added to G(Q,E).

3.4 Results

From Degree Distribution of B(Q,U), we can observe that the degree distribution of vertices in both Q and U satisfy power laws, and most of the edges of B added later in the process have their end points pointing to a low-degree node.

Theorem 7. *For the bipartite graph B(Q,U) generated after n steps, almost surely, when $n \to \infty$ the degree sequence of nodes in Q (resp. U) follows a power law distribution with exponent $\alpha = 2cq\beta/(cu(1-\beta))$*

Lemma 8. *If a node in B(Q,U) has degree g(n) at time n, with $g(n) \in \omega(logn)$, it had degree with high probability (g(n)) also at time σn, for any constant $\sigma > 0$.*

Also, we can observe that the degree distributions of the graphs G(Q,E) is heavy-tailed, and most of the nodes have degrees in $\Theta(1)$.

4 Discussion in Affiliation and Groups in Social Network

When analyzing a paper, it is important to find what problem does the work aims to solve, how the work adresses the problem, and what new insights, techniques, or system does this work contributes.

The paper for Affliations Networks (2008) points out that current social network models dont match reality as there's no densification and no diameter shrinkage, and provides analytic approach, by drawing from sociology. This work provides a model that better matches reality that is mathematically tractable and algorithmically useful.

The paper for Group Formation in Large Social Networks (2006) examines social network structure and creates a model that features decision trees to better understand how communities form and evolve. This work provides an interesting observation that the topics move before people.

The paper Structure and Evolution of Online Social Networks (2006) focuses on understanding the same problem of how communities form and evolve by measuring migration patterns and using a model of biased preferential attachment. This work provides a model that reproduces structures of two very different networks and shows that the density of network goes through 3 stages with growth.

The paper Statistical Properties of Community Structure in Large Social Info Nets (2008) focuses on the problem that We dont know much about statistical properties of network communities. This work uses five-part story of interaction graph, group interaction hypothesis, objective function to measure group-ness, clustering algorithm, and evaluation. This work provides another genarative model, possibly better than the previous ones, and points out few surprising aspects about real social networks.