Great Ideas in Computing

University of Toronto CSC196
Fall 2025

Class 20: November 14 (2025)

Announcements and Agenda

Announcements

- I have posted the third and final question for Assignment 4. Assignment 4 is due Wednesday, November 26 at 3PM.
- Our fourth and final guest presenter will be Jonathan Panuelos who will discuss "computational mechanics" which relates to the simulation of physical systems. His presentation is this coming Wednesday, November 19.
- The second and final quiz will be held Friday, November 28 in the usual tutorial room BA 2139.

Today's Agenda

- Continue Social networks

Preferential attachment models

Preferential attachnment models (also called “rich get richer” models) are probabilistic generative models explaining how various networks can be generated. Namely, after starting with some small graph, when we add a new node u, we create a number of links between u to some number m of randomly chosen existing nodes $v_1, v_2, \ldots, v_m$. The probability of choosing a v_i is proportional to the current degree of v_i.

These models have been used to help explain the structure of the web as well as social networks. Furthermore, networks generated by such a process have some nice structural properties allowing for substantially more efficient algorithms than one can obtain for arbitrary graphs.

For such models, there are both provable analytic results as well as experimental evidence on synthetic and real networks that support provable results that follow from the model. Remember, a model is just a model. A network model is a simplification of real networks; they may not account for many aspects in a real network or may not even be a good model of a real network.

Consequences for networks generated by a preferential attachment process

There are many properties, believed and sometimees proven. about preferential attachment network models that do not hold for uniformly generated random graphs (e.g., if we create random sparse graphs with constant average degree by choosing each possible edge with say probability proportional to $\frac{1}{n}$).

One of the most interesting and consequential properties is that vertex degrees satisfy a *power law distribution* in expectation. Specifically, the expectation $P(d)$ of the fraction of nodes whose degree is d is proportional to $d^{-\gamma}$ for some $\gamma \geq 1$. Such a distribution is said to have a *fat tail*.

In a uniformaly random sparse graph (with average degree d_{avg}), with high probability , the fraction of nodes having a large degree $d > d_{avg}$ is proportional to c^{-d} for some $c > 1$.

The Barabasi and Albert preferential model

Barabasi and Albert [1999] specified a particular preferential attachment model and conjectured that the vertex degrees satisfy a power law in which the fraction of nodes having degree d is proportional to d^{-3}.

They obtained $\gamma \approx 2.9$ by experiments and gave a simple heuristic argument suggesting that $\gamma = 3$. That is, $P(d)$ is proportional to d^{-3}

Bollobas et al [2001] prove a result corresponding to this conjectured power law. Namely, they show for all $d \leq n^{1/15}$ that the *expected* degree distribution is a power law distribution with $\gamma = 3$ asymptotically (with n) where n is the number of vertices.

Note: It is known that an actual realized distribution may be far from its expectation, However, for small degree values, the degree distribution is close to the expectation.

When we say that a distribution $P(d)$ is a power law distribtion this is often meant to be a "with high probability" whereas many results for networks generated by a preferential attachment process the power law is usually only in expectation.

Proven or observed properties of nodes in a social network generated by preferential attachment models

In addition to the power law phenomena, other properies of social networks have been observed. For example, a relatively large number of nodes *u* have the following properties:

- high clustering coefficient defined as : $\frac{(u,v),(u,w),(v,w) \in E}{(u,v),(u,w) \in E}$. That is, mutual friends of *u* are likely to be friends.
- high centrality ; e,g, nodes on many pairs of shortest paths.

Brautbar and Kearns refer to such nodes as “interesting indiviudals” and these individuals might be candidates for being “highly influential individuals”. Bonato et al [2015] refers to such nodes as the *elites* of a social network.

Other proven or observed properties of networks generated by preferentical attachment models

- correlation between the degree of a node *u* and the degrees of the neighbouring nodes.
- the graph has small diameter; suggesting "6 degrees of separation phenomena"
- relatively large dense subgraph "communities"'.
- rapid mixing (for random walks to approach stationary distribution)
- relatively small (almost) *dominating sets*.

For those interested, In my spring CSC303 webpage, I have posted a number of papers on elites in a social network and preferential attachment.

Some social network phenomena we want consider

We will quickly present a number of studies that illustrate the use of graph structure in obtaining information in a social-network. We will consider:

- The centrality and influence of a node.
- Detecting communities.
- Detecting the romantic relation in a Facebook network
- The importance of triadic closure and low clustering coefficient.
- The six degrees of separation phenomena

Florentine marriages and “centrality”

- Medici family connecte (by marriage)d to more families, but not by much
- More important: Medici’s lie between most pairs of families
  - shortest paths between two families: coordination, communication
  - Medici’s lie on 52% of all shortest paths; Guadagni 25%; Strozzi 10%

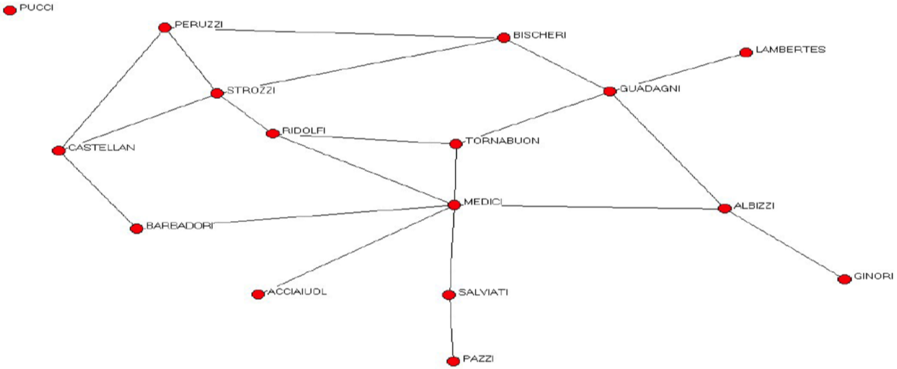

Figure: see [Jackson, Ch 1]

Example of communities and central nodes

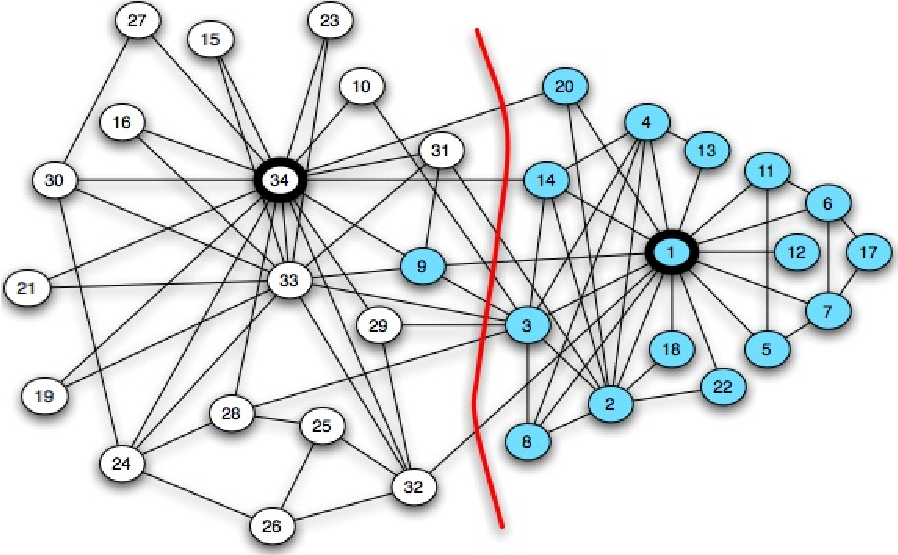

Figure: Zachary Karate Club [1977]. A *min cut* partition of the network. Note the centrality of nodes 1 and 34, the club president and the instructor.

How graph structure can reveal personal information: Detecting the romantic relation in Facebook

- There is an interesting paper by Backstrom and Kleinberg (http://arxiv.org/abs/1310.6753) on detecting “the” romantic relation in a subgraph of facebook users who specify that they are in such a relationship.
- Backstrom anbd Kleinberg construct two datasets of randomly sampled Facebook users: (i) an extended data set consisting of 1.3 million users declaring a spouse or relationship partner, each with between 50 and 2000 friends and (ii) a smaller data set extracted from neighbourhoods of the above data set (used for the more computationally demanding experimental studies).
- The main experimental results are nearly identical for both data sets.

Detecting the romantic relation (continued)

- They consider various graph strucutral features of edges, including
  1. the *embeddedness* of an edge (A, B) which is the number of mutual friends of A and B.
  2. various forms of a new *dispersion* measure of an edge (A, B) where high dispersion intuitively means that the mutual neighbours of A and B are not "well-connected" to each other in the graph without A and B.
  3. One definition of dispersion given in the paper is the number of pairs (s, t) of mutual friends of u and v such that $(s, t) \notin E$ and s, t have no common neighbours except for u and v.
- They also consider various "interaction features" including
  1. the number of photos in which both A and B appear.
  2. the number of profile views within the last 90 days.

Embeddedness and disperison example from paper

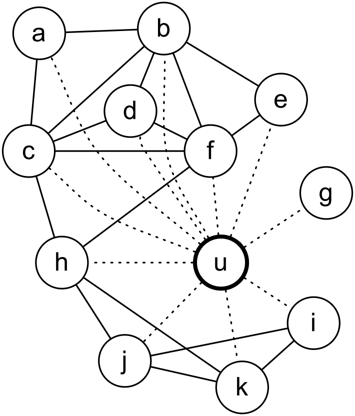

Figure 2. A synthetic example network neighborhood for a user u; the links from u to b, c, and f all have embeddedness 5 (the highest value in this neighborhood), whereas the link from u to h has an embeddedness of 4. On the other hand, nodes u and h are the unique pair of intermediaries from the nodes c and f to the nodes j and k; the u-h link has greater dispersion than the links from u to b, c, and f.

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?
- By itself, dispersion outperforms various interaction features.

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.
- By combining many features, structural and interaction, the best performance is achieved using machine learning classification algorithms based on these many features.

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.
- By combining many features, structural and interaction, the best performance is achieved using machine learning classification algorithms based on these many features.
- There are a number of other interesting observations but for me the main result is the predictive power provided by graph structure although there will generally be a limit to what can be learned solely from graph structure.

Some experimental results for the fraction of correct predictions

Recall that we argue that the fraction might be .005 when randomly choosing an edge. Do you find anything surprising?

type	embed	rec.disp.	photo	prof.view.
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (fem)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
engaged	0.179	0.446	0.442	0.391
engaged (fem)	0.171	0.399	0.386	0.401
engaged (male)	0.185	0.490	0.495	0.381
relationship	0.132	0.344	0.347	0.441
relationship (fem)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

type	max. struct.	max. inter.	all. struct.	all. inter.	comb.
all	0.506	0.415	0.531	0.560	0.705
married	0.607	0.449	0.624	0.526	0.716
engaged	0.446	0.442	0.472	0.615	0.708
relationship	0.344	0.441	0.377	0.605	0.682