Social and Information Networks

University of Toronto CSC303
Winter/Spring 2019

Week 7: February 26,28 (2020)

Announcements

- Five questions have been posted for the second asasignment. There will be a couple more questons.
- Midterm March 4 and March 6. The test will cover everything in the first six week. Not sure if we will also have a short question on this weeks material. The Wednesday, March 4 part of the test is in the usual tutorial rooms. The Friday, March 6 part of the text will be in GB248 and for those in the other tutorial, the test will take place in Haultain Building, room 403. Please take the test in the appropopriate room so that there will be plenty of space.
- Comments on the critical review assignment.
  - Due date: March 30
  - You need to find a conference or journal article that has appeared in the last 3 years; to be precise lets say, has appeared since January 1, 2017.
  - The article can be about any topic in the course.
  - You are to provide a critical review of the article as if you were on program committee or a reviewer for a journal. I will elaborate in class.

This weeks agenda

- Brief introduction to sublinear time algorithms
- Chapter 19: Influence spread in a social network

The computational challenge presented by super large networks

As we observed at the start of the course, the size of some modern networks such as the web and social networks such as Facebook are at an unprecedented scale.

The number of Web documents (e.g., distinct URLs) as reportedly indexed by Google is around 62 million pages. I have not seen a claim about the total number of links but lets say that there are perhaps a trillion or more edges in this netwok.

Facebook has roughly 2.5 billion monthly active users worldwide. The average facebook user has 155 friends which then implies about 387 billion edges. It is interesting to note that 90% of daily active users are outside USA and Canada. See https://www.omnicoreagency.com/facebook-statistics/

Linear is the new exponential

In complexity theory (e.g. in the P vs NP issue) we say (as an abstraction) that polynomial time algorithms are "efficient" and "exponntial time" is infeasible. (There are, of course, exceptions but as an abstraction this has led to invaluable fundamental insights.

As problem instances have grown, there was a common saying that "quadratic (time) is the new exponential".

But with the emergence of networks such as the web graph and the Facebook network, we might now say that "linear is the new exponential" when it comes to extracting even the most basic facts about these networks. For example, how do we even estimate the size (number of nodes and/or edges) of a giant network?

There are many facts about large networks that we would like to extract from the network. How do we compute page ranks? How do we find influential or interesting nodes" in a social network?

Sublinear time and local computation algorithms

There is a large body of research concerning sublinear time and local computation algorithms that precedes the more current interest in such algorithms in the context of specific networks like the web or large social networks.

I am providing a reasonably recent survey by Rubinfeld and Blais on the work that is not especially focused on specific networks.

I am also attaching a paper by Braubach and Kearns which is more directly relevant to our course and concerns networks that satsify power laws and more specifically, arise from preferential atachment models.

There are some notable differences between these two research "communities" but there is also a lot of similarity.

A quick comparison of the two local information and sublinear time communties

Just for the sake of this comparison, let me call the more "established area", the *graph property testing (GT)* area, and the more recent area with focus on the web and social networks, the *prefential attachment (PA)* area. This is an abuse of terminology as both areas do more than just what the name suggests. First lets consider what is common between the GT and PA areas.

What do we mean be sublinear time and local information algorithms?

These areas refer to sequential time in contrast to the area of massively parallel computation (MPC) models where we can also achieve sublinear time by distributing computation amongst a large number of processors. Although there are some connections we are focusing on sequential algorithms.

Quick comparison continued; the similarities

In general when we measure complexity, we do so as a funtion of the input/output size. For graphs $G = (V, E)$, the size of the input is usually the number of edges E. (An exception is that when the graph is presented say as an adjacency matrix, the size is n^2 where $n = |V|$.)

Given our interest in massive information and social networks, we consider sparse graphs (e.g. average constant degree) so that $|E| = O(|V|)$ and hence we will mean sublinear in n. The desired goal will be time bounds of the form $O(n^\alpha)$ with $\alpha < 1$ and in some cases maybe even $O(\log n)$ or $polylog(n)$. The literature in the GT area considers both dense and sparse graphs but again we are only considering sparse graphs.

We will always need a way to access these massive graphs.

Given that almost any optimal algorithm for a graph property (respectively, for any function) will depend on the entire graph (resp. the entire input) , we will have to settle for approximations. Furthermore, we will need to sample the graph so as to avoid having to consider all nodes and edges.

Local computation/information algorithms

Most (if not all) of the research in both the GT and PA areas are with respect to local computation (as mainly refered to in the GT area) or local information (as mainly refered to in the PA area) algorithms. There are sometimes differences in the way these terms are used but we will use the following informal idea.

Local information algorithms are sequential algorithms where the network topology is initially unknown and is revealed only within a local neighborhood of vertices, and when information is revealed, this information is then irrevocably added to the unknown network.

To make this more precise we will present the *Jump and Crawl* computation model as used in the PA area.

The Jump and Crawl Model

The Jump and Crawl model assumes access to a uniformly revealed random nodes (the Jump operation), and the ability to examine any and all neighbours of any revealed vertex (the Crawl operation).

In the web graph, Crawl is going to a link from a page, and in Facebook, crawl going to a friends profile. Jump is like the scaled version of page rank and, in Facebook, like a generalization of a friends finder option to find someone with a given profile.

Quick comparison continued; the differences

In mentioning the differences, I am only refering to what seems to be the emphasis in these areas; it is also apparent that any differences between the models are not fundamental inherent differences.

In the GT area, the vertex names are known and impliclty so is the size $|V|$ of the network. (Again, I am restricting attention to the sparse GT model.) See, however, a recent paper by Goldreich [2018].

In the GT area, algorithms explore adjacent edges without revealing all adjacencies at once.

In the GT area, the focus is often on global graph properties (e,.g. average degree) whereas in the PA area the emphasis is often on extremal properties (vertices of maximum degree).

However, the main difference for me is that the GT area is more focused on worst case (over all graphs) rather than graphs which enjoy properties that result say from a preferential attachment model as in the PA work. (But here results are often contrasted with results for general graphs.)

A brief introduction to results in the PA area

The area of sublinear time and local information algorithms is quite extensive and some results are quite technical. For now we will mainly just give an overcview of results in the Brautbar and Kearns paper, a paper by Bonato et al [2015], plus some obervations in an unpublished paper by Ben-Eliezer, Eden, Fotakis and Oren [2020].

I hope to return to this topic later in the term.

Brautbar and Kearns are using the prefential attachment model in a paper by Barabasi and Albert [1999] which is slightly different from the model as discussedd in the EK text. In their model, the process starts with a fixed number (say n_0) of vertices and then vertices are added to the graph one at a time and joined to n_0 earlier vertices, selected with probabilities proportional to their degrees.

The power law for the Barabasi and Albert Model

Barabasi and Albert suggested that after many steps the proportion $P(d)$ of vertices with degree d should obey a power law $P(d)$ proprortional to $d^{-\gamma}$.

They obtained $\gamma \approx 2.9$ by experiments and gave a simple heuristic argument suggesting that $\gamma = 3$.

Bollobas et al [2001] provide a provable result corresoponding 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 expectation.

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

End of Wednesay, February 26 lecture

We ended the lecture mentioning the preferential attachnment mlodel of Barabasi and Albers and the fact that (in expectation) the node degrees $P(d)$ satisfy a power law with exponent 3; i.e. $P(d)$ proprtional to d^{-3}.

Observed properties of social networks generated by such preferential attachment models

In addition to this power law phenomena, other properies of social networks have been obseerved such as :

- Nodes having
  - high degree
  - high clustering coefficient
  - high centrality
  - These are what Brautbar and Kearns call sets of "interesting individuals' and might be candidates for being "highly influential individuals". Bonato et al [2015] refers to such nodes as the *elites* of a social network.
- Plus other properties such as
  - small diameter
  - relatively large dense subgraph communities.
  - rapid mixing (for random walks to approach stationary distribution)
  - relatively small (almost) *dominating sets* .

(Almost) dominating sets and the MGEO-P preferential attachment model.

A dominating set in a graph $G = (V, E)$ is a subset $S \subset V$ such that every $v \in V$ is adjacent to at least one vertex $s \in S$. It is NP-hard to determine a dominating set within a $\log n$ factor of a minimum size dominating set. One can relax this to call a set almost dominating if "almost all" are adjacent to at least one vertex $s \in S$.

Bonato et al [2015] consider a preferential attachment model called MGEO-P. This model is a geometric model where metric distances reflect homophily. They show that in this model, networks have sublinear dominating sets and they empirically verify the presense of a sublinear dominating set consisting of high degree vertices within a large subset (hundreds of millions) of the Facebook network.

Ben-Elizer et al consider the micro blogging site Tumblr. They call a user engaged if they follow at least 10 and find that 35% ($\approx$ 88 million) of users are engaged. They observe that 99.3% (resp 98.2%, 94.6%) of engaged users follow at least one of the top 1% (resp .1%, .01%) of highest degree users.

Elites in a social network

Bonato et al conclude their paper with the following comment:
A different approach to detecting elites is to search for them within a minimum size dominating set, as these sets reach the entire network. Further, if minimum size dominating sets have much smaller order than the network (as we postulate), then that reduces the computational costs of finding elites.

This theme of utliizing small dominating sets is pursued in the unpublished paper by Ben-Elizer et al. Their observations about small dominating sets in the Tumblr network reflect what is known as Price's square root law.

Price's law is due to Derek J. de Solla Price who (in 1965) gave perhaps the first mathematical rich get richer scale free network model. He was interested in explaining how citation networks grow. Price's Law states that half of the work that a group does is completed by the square root of the number of people in the group.

See Avin et al [2018] paper entitled "Elites in social networks: An axiomatic approach to power balance and Price's square root law".

Paragraph in Avin et al

In economy, recent reports show that the gap between the richest people and the masses keeps increasing, and that decreasingly fewer people amass more and more wealth [10, 11]. Claims like "The top 10 percent no longer takes in one-third of our income – it now takes half," made by former president Obama [12] when addressing the issue, are interpreted as implying that the economic and political elites become increasingly more greedy and overbearing. Such claims are often used in order to criticize governments and regulatory financial institutions for neglecting to cope with this disturbing development. The question raised by us is: can society help it, or is this phenomenon an unavoidable by-product of some inherent natural properties of society? We claim that in fact, one can predict the shrinkage of elite size over time (as a fraction of the entire society size) based on the very nature of social elites. In particular, in our model, such shrinkage is the natural result of a combination of two facts: First, *society grows*, and second, *elites are much better connected* than peripheries. Combining these facts implies that the fraction of the total population size comprising dense elites will decrease as the population grows with time. And this is what we call Price's square root law in networks and, in particular, Price's law for elites in social networks.

And after this little bit of political science, we return to our interest in subinear time algorithms for massive information and social newtorks

Highlights of some resuts in Brautbar and Kearns

Brautbar and Kearns prove (using the Jump and Crawl model) a number of interesting (positive and negative) results contrasting what can be shown for general networks vs what holds for power law, and preferential attachment networks. We consider their results for finding a high degree vertex with a provable appoximation.

I will only state results (using slides by Brautbar) without being precise about some definitions and assumptions. The paper gives three algorithms (for arbitrary networks , for power law networks, and for networks generated by the PA attachment process). We will briefly present their results for finding a high degree vertex.

An $c(n)$ approximation algorithms for max degree (and similarly for any maximization problem) returns a vertex v such that $deg(v^*) \leq c(n)deg(v)$ where v^* is a vertex of maximum degree.

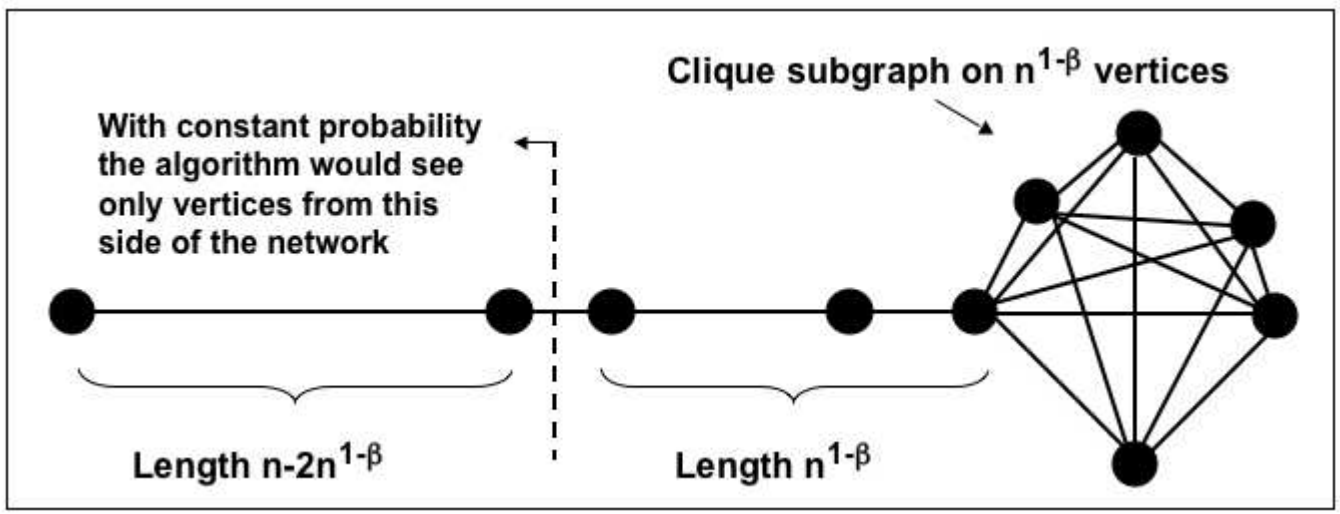
Theorem for approximate max degree vertex in an arbitrary network

Let G be an arbitrary network and let v^* be a vertex in G of maximum degree say d^*. Then for any $0 < \beta \leq 1$, there is a relatively simple algorithm using $\tilde{O}(n^{\beta})$ Jump and Crawl operations that w.h.p (with high probability) returns a vertex v such that $d^* \leq n^{1-\beta} deg(v)$.

Here is the algorithm:
If $d^* < n^{1-\beta}$ then any vertex will suffice.
Else
  For $\tilde{O}(\frac{n}{d^*})$ trials
    Jump to a random vertex
  % The claim is that with this many queries, a neighbor of v^* will be found.
    If $deg(v) \geq \frac{d^*}{n^{1-\beta}}$ we are done
    Else Crawl all of $v's$ neighbors to see if one has high degree.
  End For

A matching negative result

There is a matching lower bound. In the graph below, with constant probability, n^{β} jump operations will still leave us a distance $n^{1-\beta}$ away from the clique part of the network.



Theorem for approximate max degree vertex in a power law network

Let G be a power new with expomnent $\gamma > 2$. Let $0 < \beta < \frac{\gamma-1}{\gamma}$. Then there exists an algorithm using $\tilde{O}(n^{\beta})$ Jump and Crawl operations that w.h.p finds a vertex with expectded approximation ratio $O(n^{\frac{1}{\gamma}-\frac{\beta}{\gamma-1}})$.

The algorithm simply jumps to $\tilde{O}(n^{\beta})$ random nodes and takes the vertex of highest degree.

The approximation guarantee relies on two properties of power law networks with exponent $\gamma > 2$. Namely

1. The highest degree is $O(n^{\frac{1}{\gamma}})$ and
2. The probability of randomly sampling a node of degree at least $n^{(\frac{\beta}{\gamma-1})}$ is $\Theta(n^{-\beta})$.

Theorem for approximate max degree vertex in the PA model of Barabasi and Albert

Let $0 < \beta < \frac{1}{11}$ and Let G be generated by the PA model of Barabasi and Albert. Then there is an algorithm using $\tilde{O}(n^{\beta}$ Crawl operations that finds a vertex with degree approximating the maximum degree with expected approximation ratio $O(n^{\frac{1}{2}-\beta})$.

The algorithm runs n^{β} *lazy random walks* from an arbitrary vertex and takes the termination vertex (of the random walk) with highestb degree.

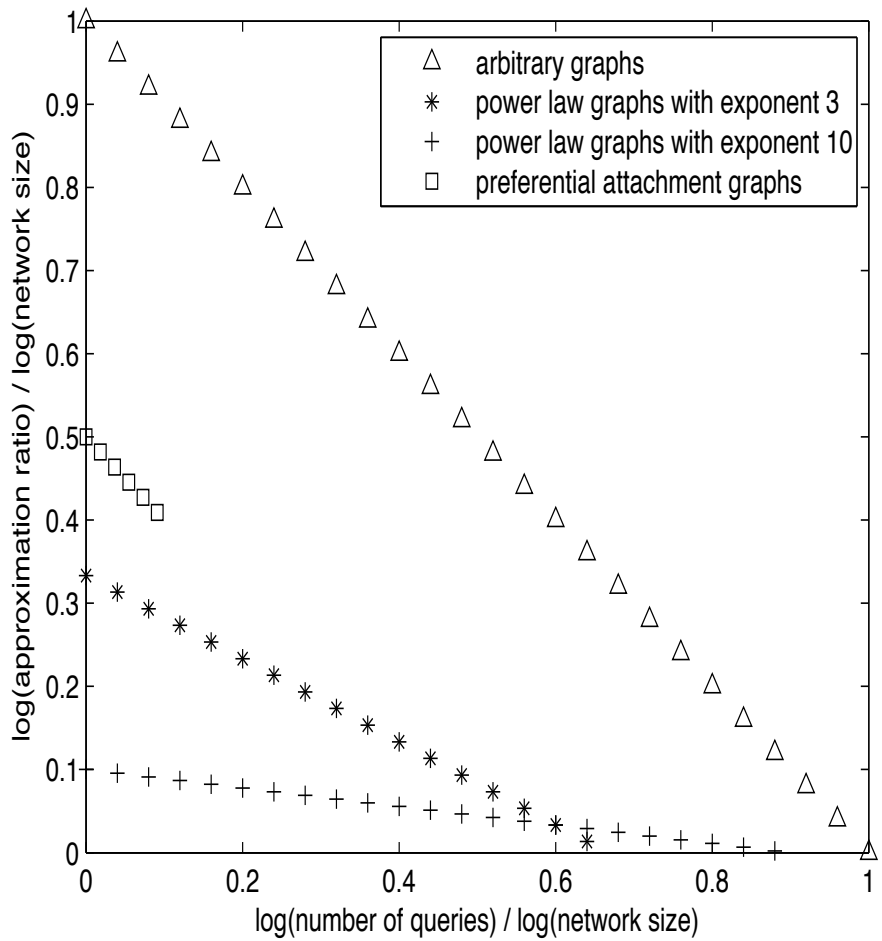
A lazy random walk is the same as the random walk in scaled page rank withing scaling factor $\frac{1}{2}$. That is, with probability $1/2$ it goes to a neighbour chosen uniformly at random and with probability $1/2$ it goes to a uniformly chosen random network node.

The approximation relies on the following facts:

1. A lazy random walk has a unique stationary distribution.
2. In this PA model, the mixing time of the lazy random walk is $\tilde{O}(\log n)$.

The comparison for finding a vertex whose degree is a good approximation to the maximum degree

The plots represent the approximation n^{δ} guarantee as a function of the number n^{β} of Jump and Crawl operations. These are log-log plots so the figure is plotting ithe exponent δ as a function of the exponent β.



End (for now) of discussion of sublinear time and local information algorithms

Brautbar and Kearns also consider the approximation computation of a vertex having both high degree and a high clustering coefficient.

Other network properties have been studied including searching for the root in a PA process, and the approximation of dominating sets, page rank, and the approximate computation of "highly influential nodes".

Influence spread is a basic issue in social networks and is similar to issues regarding disease contagion. This bring us back to the text and, in particular, Chapter 19.

End of Friday, February 28 lecture