Lab 1: The Most Popular Node

"What exactly is the importance of the highest degree node in a graph?"

Part I: Graph Structures and Graph Metrics

Graph Terminology Definitions

Social Network Graph $G = (V, E)$

- V: set of vertices = points in the graph = individuals in the population
- $n = |V|$: size of graph = number of nodes in the graph = size of population
- E: set of (undirected) edges = lines between points = mutual friendships between individuals

Node Degree $d(v)$
$d(v)$: number of edges connected to v = how many friends v has

Maximum Degree $d_{max}(G), v_{max}$

- v_{max}: a node of highest degree in G = the most popular individual in G (or possibly one from a set of maximally popular individuals)
- $d_{max}(G)$: the maximum degree of any node in G = the total number of friends v_{max} has

Minimum Degree $d_{min}(G), v_{min}$

Degree Distribution $D(G) = \{d(v) : \forall v \in V\}$
$D(G)$: the set of degrees of nodes in G = the distribution of numbers of friends across individuals in the data set.

Distance Between Nodes $\delta(a, b)$
$\delta(a, b)$: the number of edges on the shortest path connecting a and b in the graph = if a starts a meme which passes from friend to friend along the social network, $\delta(a, b)$ is the fewest possible steps for the meme to reach b.

Diameter of G
$diameter(G)$: the maximum value for $\delta(x, y)$ for any pair $x, y \in V$ = the "longest shortest path" in G = the distance between the two most separated individuals in the population.

Comprehension Checks:

1) If $n = 4$, what's the maximum number of edges G can have? In general, if you don't know d_{max}, what's the maximum number of edges G can have in terms of n? What if you *do* know d_{max}?

2) What does the population look like if $d_{min} = d_{max}(G) = (n-1)$?

3) In a graph with 4 individuals, is it possible for $D(G) = \{3, 3, 3, 2\}$? What about $\{3, 2, 2, 1\}$? What must be true about n for it to be possible to have $d_{min}(G) = d_{max}(G) = (n-2)$?

4) If I remove an individual from G, at most how many elements of $D(G)$ can change? At least how many elements of $D(G)$ *must* change? By how much will these elements change?

5) What effect does removing v_{max} from the graph have on the diameter of the graph? Specifically, find a graph G which maximizes this effect.

6) Say I want to count the number of triangles in G (ie the number of instances where a set of three people are all friends with each other). This is used to compute the 'clustering coefficient' of G, ie a social network tends to be more clustered when everyone's friends are also friends with each other. What's the maximum number of triangles v_{max} can participate in? How many triangles are there in G if $d_{min} = d_{max} = n$?

Part II: How does v_{max} affect our ability to release information about the graph without endangering individual privacy?

Differential Privacy Basics:

Neighboring Datasets $D1, D2$

- n: the number of individuals in a population
- $D1$: a dataset, a collection of data about the population of n individuals
- $D2$: a 'neighboring' dataset, equal to $D1$ except that an arbitrary individual's data has either been added or removed. $D2$ includes either $(n-1)$ or $(n+1)$ individuals.
- *bob*: a convenient method of referring to the individual differing between $D1$ and $D2$.
- $G1, G2$: two neighboring social network graphs, differing in one individual; $G2$ is produced by either removing or adding one node to $G1$.

ϵ-Differential Privacy

A (randomized) query Q is ϵ-differentially private if $\forall$ pairs of neighboring datasets $D1, D2$:

$$\frac{P(Q(D_1) = A)}{P(Q(D_2) = A)} \leq e^{\epsilon}$$

Intuitively: If given the privatized answer A, it's fairly equally likely to have come from $D1$ or D, then we can't tell whether or not *bob* was in the data set, and thus *bob*'s privacy is protected. The value of ϵ is chosen by the person developing the privatized query; it characterizes the trade-off between privacy and accuracy.

Function Sensitivity

- Function $F(D) = \{a_1, a_2, ...a_n\}$: In this case, F is any function that takes as input a dataset D and returns a set of numerical values. For instance, if D_{Hair} is a list of summer school students along with their hair colors, then $F_{Brown}(D_H)$ could return the count of students with brown hair. Or, $F_{Hist}(D_H)$ could return a histogram of the students with relation to their hair color. For F's that return a set of values, we'll use the notation that $F(D)_i = a_i$
- Sensitivity $\Delta(F)$: First, if $D_1, D2$ are neighboring datasets, then we'll say $F(D1) - F(D2) = \sum_{i=1}^{n} |F(D1)_i - F(D2)_i|$. Then $\Delta(F) = max_{D1,D2}\{F(D1) - F(D2)\}$; the sensitivity of F is the maximum possible difference $F(D1) - F(D2)$ across *any* two neighboring datasets $D1, D2$. It is the largest impact removing or adding one individual can have on the value of F.

The Sensitivity Method
If the sensitivity of query $F = \Delta(F)$, then we can create a randomized, ϵ-differential private query result $Q(D)$ by taking $F(D)$ and adding laplacian noise of standard deviation $\frac{\Delta(F)}{\epsilon}$. Intuitively, this noise *covers* the gap between $D1$ and $D2$ for any choice of $D1, D2$.

Comprehension Checks:

1) What is the sensitivity of F_{Brown} as defined above?

2) What is the sensitivity of F_{Hair}?

3) Say we form a differentially private Q_{Brown} using the sensitivity method. What sort of values might we see? Is it possible for $Q_{Brown}(D1) < 0$? Is it possible for $Q_{Brown}(D2) > n$? What can we do to produce more realistic answers without reducing privacy?

4) Let's say instead of F_{Brown} we have F_{Bear}, a list of summer school students who still sleep with teddy bears. Say that in a school of 25 students, $F_{Bear} = 6$. An attacker may have partial knowledge of a dataset; Eve knows that 5 students definitely sleep with bears and 19 students definitely don't, but she's uncertain about you. What value of ϵ are you comfortable with? If $\epsilon = ln(2)$, $D1$ = the 25 students, $D2 = D1$ - you, and it turns out privatized A we given to Eve is equal to 6.5... is that ok with you? What value of ϵ would you prefer? What would be the standard deviation of the noise you'd add to F_{bear}?

5) Given D_{shoes} = a list of some people along with the number of pairs of shoes they own. Remember that sensitivity is the maximum difference between $F(D1)$ and $F(D2)$ over *any* choice of $D1, D2$. Say $F_{total}(D_{shoes})$ is the total number of shoes owned by everyone in the list. What's the sensitivity of F_{total}? If $F_{avg}(D_{shoes})$ is the average number of shoes owned by each person in the list, what's its sensitivity? Say $F_3(D_{shoes})$ returns 1 if someone in the list owns three pairs of shoes, and returns 0 otherwise. What's its sensitivity?

6) (tricky) If, as in the case of F_{Brown}, I only want to cover a gap of one person, why not just pick a *noise* value with equal probability from $\{-1, 0, 1\}$ and return $Q_{Brown} = F_{Brown} + noise$?

Interesting Questions

1) Remember that $G1, G2$ are neighboring graphs if $G2$ is produced by adding or removing one node from $G1$. What is the sensitivity of $diameter(G)$?

2) A brief definition: Say G is a social network representing high school students and teachers. There are no friendships between students and teachers, so G has two completely separate groups of nodes. We say that G has two 'connected components'. If $connected-comp(G)$ returns the count of connected components in G, what is it's sensitivity? What if we know that for any possible graph Gi that our query might run on, $d_{max}(Gi) \leq DMAX$?

3) What is the sensitivity of the degree distribution $D(G)$ if DMAX is known?

4) What is the sensitivity of $triangle-count(G)$ if DMAX is known? When is it possible to release a reasonably accurate triangle count?

Part III: How does v_{max} affect our analysis of the structure of the graph?

Graph Sampling/Estimation Basics

Graph Sampling Methods:

Studying an intractably large social network graph requires subsampling the nodes and edges of the graph in order to form a smaller, useful sample graph. Different techniques can be used to do this, and each technique has unique properties. Three of these techniques are listed below.

- Node Sampling: Nodes are selected randomly and added to the graph along with their associated edges. Intuitively: Individuals are randomly selected from the population and are added to the sample graph along with all of their friends (but not their friends' friends).
- Edge Sampling: Edges are selected randomly and added to the sample graph. Intuitively: Friendships are chosen at random, and their participants are added to the graph.
- Topology Sampling: A subgraph is chosen by a finite random walk along the graph. Intuitively: The sample graph is formed by starting with a randomly chosen individual, traveling to one of her friends (and adding him to the graph), then traveling to one of his friends (and adding him to the graph) and so on for a set number of steps.

Graph Estimation/Modeling and Graph Metrics:

One technique used in studying social networks is to attempt to model real networks with artificially built/grown networks; this produces a larger, more maliable, (and potentially more private) dataset with which to work. To determine whether an artificially built graph is sufficiently similar to the real networks it is modeling, a set of statistics called 'graph metrics' are compared between the two. These metrics are also used to determine if a sample graph accurately represents the larger graph it was selected from. Below is a list of commonly used metrics.

- $Diameter(G)$
- degree distribution $D(G)$
- hop plot: $HopPlot(G) = \{\delta(a,b)|a,b \in V\}$ = the distribution of shortest path lengths between nodes in the graph.
- $TriangleCount(G)$

- $TrianglePartication(a)$: The number of triangles a specific node a participates in. One can also look at the distribution of triangle participation levels across nodes in the network.
- $Density(G)$: the ratio $\frac{|E|}{|N|}$

Interesting Questions

1) How would you expect v_{max} to affect the values of each of the graph metrics?

2) For both node and edge sampling, in a graph G with n members, given node v_1 of degree $d(v_1)$, what is the probability that v_1 will be included in the sampled graph? What is the probability that (if v_1 is selected) all of v_1's friends will also be included in the sampled graph? What are the biases of node and edge sampling? Which graph metrics do they affect? How, generally, would you expect topological sampling to behave in comparison?

3) Say a topological sampling starts on node a, and G is a complete graph of size n (it has the maximum possible number of edges). If the sampling runs for 3 steps, what's the probability that a's triangle participation in the resulting graph will be at least 1? Analyze the relationship between triangle-participation and topological sampling as far as your desire and your background in probability allow.

4) How can the removal of v_{max} affect the triangle-participation for a different node a? How might it affect the distribution of triangle participations?

5) Take the hop plot, and for each sampling technique find a pathological graph G where the hop plot histogram is likely to vary wildly between sampling outcomes.

6) What is the relationship between the robustness (ie, resilience to flaws in the data or sampling) of the graph metrics used for graph characterization and the privatizability of those metrics?

Conclusion: Things to Think About

How much of what we see as the meaning of a social network graph is affected by the popularity level of the most popular people? What does this imply for graph dynamics (as popularity relationships change)? What does it imply for attempts to preserve the privacy of everyone in the graph?

There are cases where some members of a social network have an artificially high popularity (say the facebook page made for the university mascot); how can this change our perception of the nature of the graph? What about graphs with two classes of members, popular and unpopular (say friendly cheerleaders and football stars who accept all friendship requests of admiring strangers, and average students who tend to friend only real-life companions); how can this disparity in node degrees influence our perception of the graph structure as characterized by graph metrics? Might it lead to any inaccurate assumptions about the social network?

Can you think of any graph properties which are relatively insensitive to high degree nodes?