Differentially Private Partitioned Graph-mining

Christine Task, Chris Clifton
Computer Science, Purdue University

Background

Differential Privacy [1]:

A data-set D *neighbors* a data-set D', if they differ in one individual. Given a data-mining query Q on a data-set D, ε-differential privacy requires adding sufficient noise so that guessing whether original query was Q(D) or Q(D'), on any neighbor D', is unlikely. Thus, an attacker seeking information about Alice, and given the privatized result R = [Q(D) + *noise*], will not be able to discover from R whether Alice was even involved in the query, much less what her data was. Formally, R satisfies ε-differential privacy if, for any neighbors D, D':

$$\frac{P(R(D) = a)}{P(R(D') = a)} \leq e^{\varepsilon}$$

Global Sensitivity:

The global sensitivity Q_Δ of a query Q is the largest effect adding or removing one individual can have on the query result Q(D), over all possible choices of D. Intuitively, this is the "gap" that must be obfuscated in order to achieve ε-differential privacy. In general, R = [Q(D) + *Lap*(Q_Δ)] is ε-differentially private, where *Lap*(Q_Δ) is a random noise value drawn from the Laplacian distribution with mean 0 and magnitude *Lap*(Q_Δ). However, privatizing queries with high sensitivity may require noise levels which obliterate the results.

Differentially-Private Graph-mining:

Social network analysis is an obvious application for the strict guarantees of ε-differentially private techniques: interesting graphs may contain sensitive data, and the wide availability of public social networks gives attackers considerable access to outside information. Differential privacy ensures that regardless of what an attacker knows, he cannot use privatized results to learn about individuals.

Graphs: Unbounded Global Sensitivity!

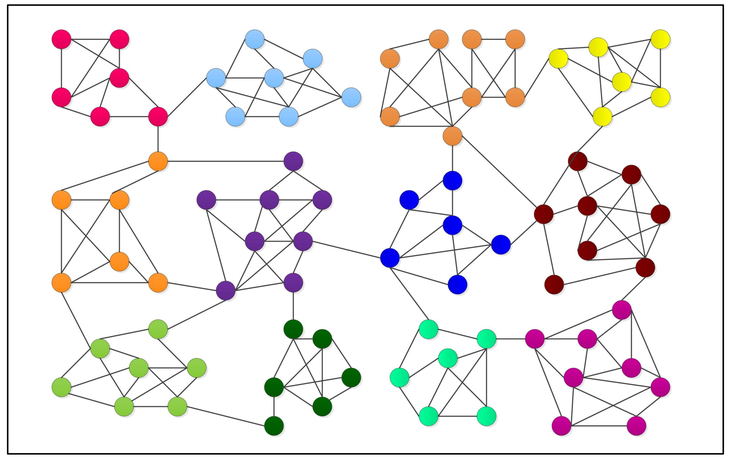
Unfortunately, due to the large impact removing or adding a node can have on most graph metrics, graph-mining queries often have unbounded sensitivity. Consider triangle counts: One individual can be involved in at most ((n-1) choose 2) = (n-1)(n-2) /2 distinct triangles in a graph of size n, so removing her reduces the count by that amount. But, global sensitivity is the maximum over the *entire* domain space of inputs, including all possible, arbitrarily large, graphs G. Since n is unbounded, $Tri_\Delta(G) \rightarrow \infty$

Recent Approaches Reduce Privacy:

Recent approaches avoid using global sensitivity to achieve privacy, but fail to provide ε-differential privacy in (very small) δ percent of cases[2]. This may not satisfy legal requirements for data protection.

References:
[1] C. Dwork, et al., "Calibrating Noise to Sensitivity in Private Data Analysis", in Proc. TCC, 2006
[2] Vishesh Karwa, et al. "Private Analysis of Graph Structure, " In Proc.VLDB, 2011.

Proposed Special Case: Private Partitioned Graph-mining



Example: How many triangles are in each partition? What's the distribution of triangle counts across partitions?

Motivation: Why Partition?

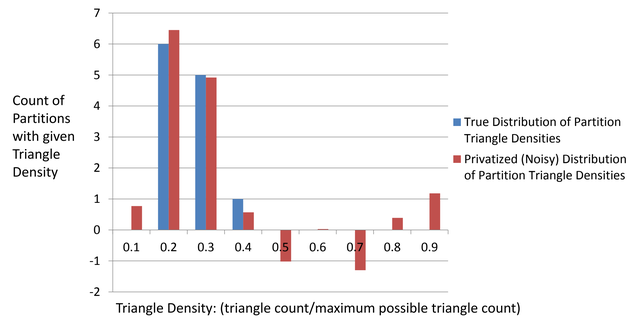
- More Meaningful: In social networks, studying patterns across social clusters may be more meaningful than treating the entire graph as a single entity.
- More Descriptive: Releasing a privatized distribution of patterns in the graph may provide more information about the graph structure than a single graph metric value for the graph as a whole.
- Privatizable: An individual can only affect the graph metric value for their own partition. This bounds the sensitivity of the function and enables us to protect the privacy of all individuals.

Differentially-Private Partitioned Graph-mining:

1. Partition graph using existing node attributes
   Examples of useful clustering attributes: dorm, language, hometown, interest-group membership
2. Calculate the desired graph metric for every partition
   Interesting metrics: triangle count (indicative of social cohesion), diameter, average number of friends
3. Normalize results, mapping them to [0,1] range
   This mapping ensures the range of results is independent of the data; it can be complex and constructed to limit precision loss.
4. Chart distribution of results as a histogram
   The histogram allows us to release meaningful data about the distribution of results across partitions, with low sensitivity.
5. Add noise of magnitude 2 to histogram values
   If one individual is added or removed, at most one partition may shift to a different histogram bucket (Global Sensitivity = 2).
6. Release privatized histogram

Privacy Guarantees, Applications, Next Steps

Example of Privatized Data:



Privacy Guarantees:

- Partition Attributes Remain Private:
  Because a person's partition membership is based on attributes independent of the graph structure (so adding or removing a neighbor does not affect an individual's partition assignment), partitioning schemes have zero sensitivity and will not leak information about individuals.
- Released Histogram Satisfies Differential Privacy:
  The histogram has a sensitivity of 2, allowing ε-differential privacy with relatively little added noise.

Example Data, Interesting Questions:

- 2005 University Facebook Data-set:
  - Partitioning attributes: university, dorm, major, gender, faculty/student
  - Do co-ed dorms tend to form more or less cohesive social networks?
  - Do humanities or science majors have more friends on average?
  - What's the average degree of separation between faculty members?
  - Are friend groups in science/engineering disciplines split by gender?

Next Steps:

This work is at a very early stage. Here's what we plan to do next:

- Implement algorithm, gather experimental data
- Develop more sophisticated normalization mappings
- Release more detailed results, ex: nested histograms
- Collaborate with social science researchers
- Develop a differentially-private partitioning algorithm to replace step 1 on non-partitioned graphs