Preserving the Privacy in Online Social Networks Using Enhanced Clustering Algorithms

Rumman Ahmed

Outline

- Introduction
- Related Work
- Problem Statement
- Experimental Results
- Conclusion

Introduction

- The role of social networks for the modern population
- Public and private data sharing by users
- Collection of data from social networks
- Protecting the anonymity of a social network

Introduction - Public Data Sharing

Privacy Settings and Tools

Who can see my stuff?	Who can see your future posts?	Only Me	Edit
	Review all your posts and things you're tagged in		Use Activity Log
	Limit the audience for posts you've shared with friends of friends or Public?		Limit Past Posts
Who can contact me?	Who can send you friend requests?	Friends of friends	Edit
Who can look me up?	Who can look you up using the email address you provided?	Friends of friends	Edit
	Who can look you up using the phone number you provided?	Friends of friends	Edit
	Do you want search engines outside of Facebook to link to your Profile?	No	Edit

Introduction

- The role of social networks for the modern population
- Public and private data sharing by users
- Collection of data from social networks
- Protecting the anonymity of a social network

Related Work

- Sequentially clustering to anonymize a network
- The k-member clustering problem
- Data and structural k-anonymity in social networks

Problem Statement

- A social network can be represented as a graph
- Clustering a social network..the issue at hand

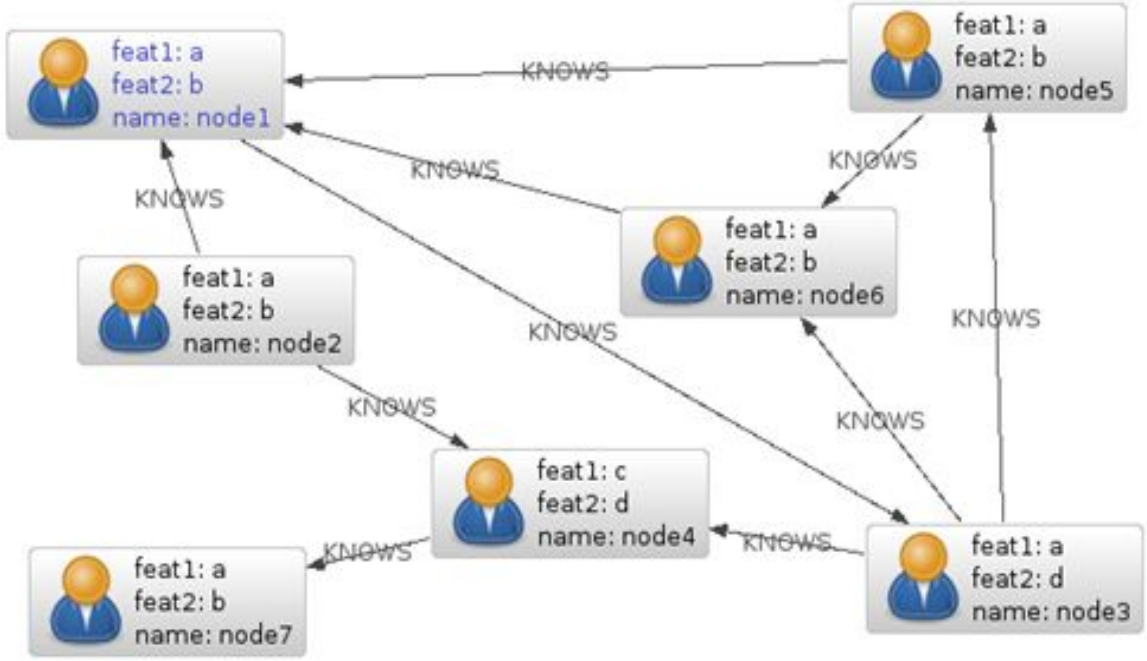


Figure 1. A social network as a graph

Problem Statement

- Preventing single user clusters
- Keeping information loss at a minimum
- Increasing degree of anonymity

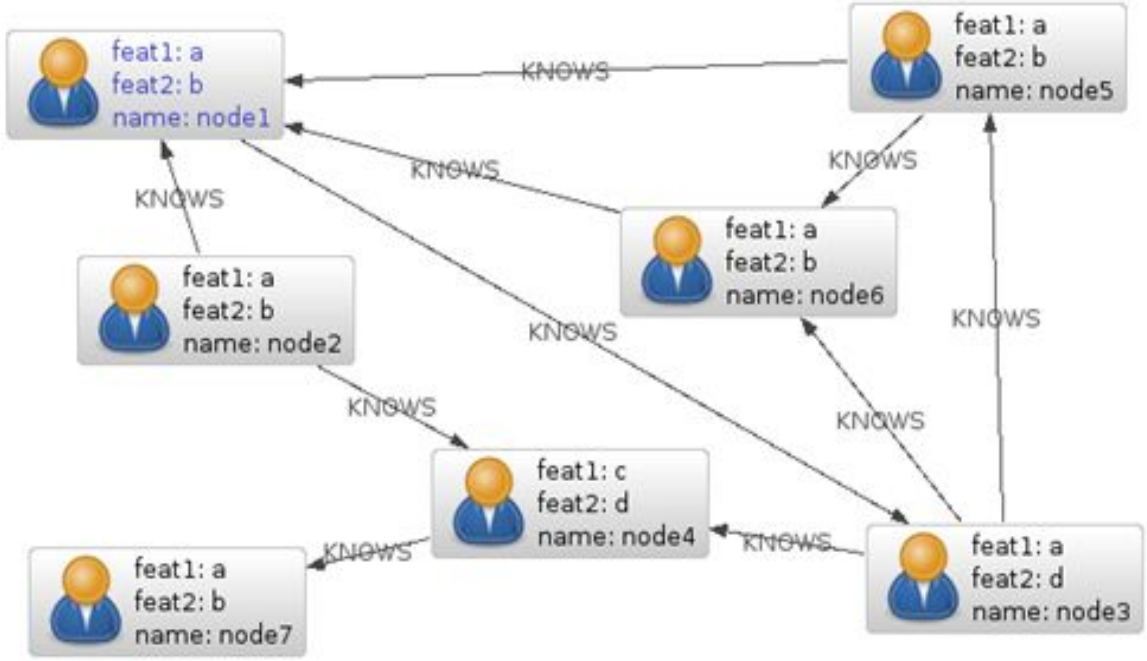


Figure 1. A social network as a graph

Problem Statement - The Algorithm in Use

Algorithm 1: (

1. Determine cl
2. Calculate ini
3. Sort points b
4. Assign point
   with all point
5. Compute cur
6. For each poi
7. Sort points b
   the best poss
8. For each poi
   a. For e
      i.
   b. If the
9. If no more tr

Algorithm 1: Generating Equal-Sized Clusters

1. Determine cluster size
2. Calculate initial clusters with built-in original clustering algorithm
3. Sort points by the difference of their nearest cluster to the farthest cluster
4. Assign points to nearest cluster until the cluster size is reached. Continue with all points
5. Compute current cluster means
6. For each point, compute the distances to the cluster means
7. Sort points based on the difference between the current assignment and the best possible alternate assignment.
8. For each point by priority:
   a. For each other cluster:
      i. If there is an element wanting to leave the other cluster swap the two elements if it's an improvement, without violating cluster max size
   b. If the element was not changed, add to outgoing transfer list.
9. If no more transfers were done (or max iteration threshold was reached),

Experimental Results

- Traditional clustering algorithms compared to our enhanced algorithms
- Metrics
  - Information Loss
  - Degree of Anonymization
  - Running Time

Experimental Results

- Experiment Details
  - Dataset: 5000 Yelp users
  - Enhanced versions of K-Means, Mean-Shift, and Affinity Propagation
  - Implemented in Python

Experimental Results - Tables

Table 1: K-Means, 5000 users

K (number of clusters)	Traditional			Enhanced		
	Information Loss	Degree of Anonymization	Running Time (seconds)	Information Loss	Degree of Anonymization	Running Time (seconds)
5	0.64%	0.0016	5	0.82%	0.001	215
10	0.43%	0.0004	6	0.62%	0.002	152
25	0.27%	0.0002	6	0.47%	0.004	725
50	0.17%	0.0002	8	0.33%	0.01	711
100	0.12%	0.0002	10	0.3%	0.02	950

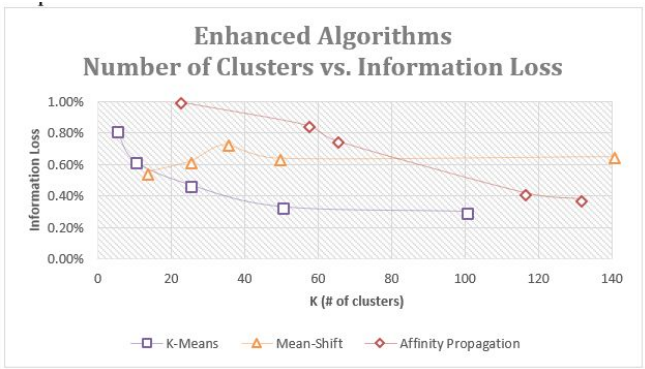
Table 2: Mean-Shift, 5000 users

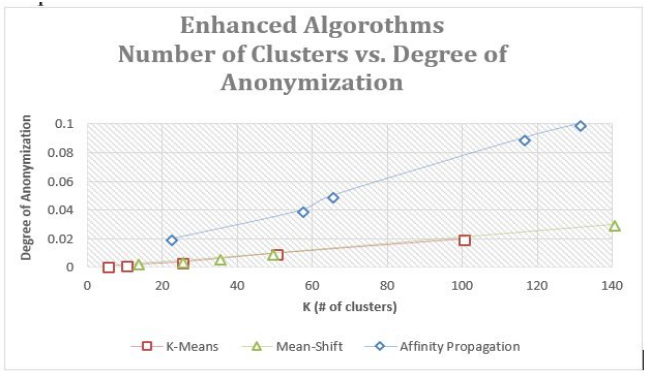
K (number of clusters)	Traditional			Enhanced		
	Information Loss	Degree of Anonymization	Running Time (seconds)	Information Loss	Degree of Anonymization	Running Time (seconds)
13 (q 0.9)	1.1%	0.0002	35	0.55%	0.003	366
25 (q 0.5)	0.85%	0.0002	50	0.62%	0.005	564
35	0.86%	0.0002	67	0.73%	0.007	777
49 (q 0.2)	0.52%	0.0002	78	0.64%	0.01	772
140 (q 0.05)	0.21%	0.0002	104	0.65%	0.03	1656

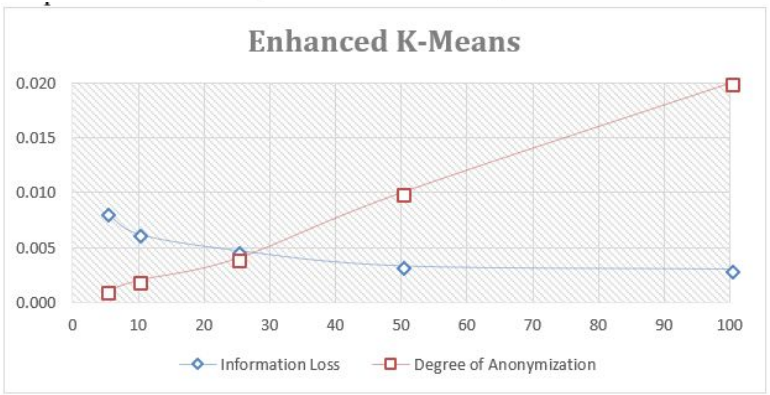
Table 3: Affinity Propagation, 5000 users

K (number of clusters)	Traditional			Enhanced		
	Information Loss	Degree of Anonymization	Running Time (seconds)	Information Loss	Degree of Anonymization	Running Time (seconds)
22 (d 0.99)	0.75%	0.0002	9	1.0%	0.02	112
57 (d 0.9)	0.25%	0.0002	13	0.85%	0.04	156
65 (d 0.85)	0.26%	0.0002	13	0.78%	0.05	158
116 (d 0.6)	0.25%	0.0002	15	0.42%	0.09	245
131 (d 0.5)	0.26%	0.0002	16	0.38%	0.1	263

Experimental Results - Graphs







Result Analysis

- Information loss kept at a minimum
- Increase of the degree of anonymization
- Running time comparisons, enhanced algorithms impractical for large data

Conclusion

- Importance of privacy and anonymity in modern time
- Current solution of clustering has a drawback
- New proposed solution shows promising results and a stepping stone for the anonymity problem

References

1. T. A. Pradesh, Sequential Clustering for Anonymizing Social Networks. *International Journal of Information & Computation Technology*, 2014.
2. A. Campan og T. M. Truta, Data and Structural k-Anonymity in Social Networks. 2008.
3. J. Byun, A. Kamra, E. Bertino, and N. Li. (2007) “Efficient k-Anonymization Using Clustering Techniques,” In: Kotagiri R., Krishna P.R., Mohania M., Nantajeewarawat E. (eds) Advances in Databases: Concepts, Systems and Applications. DASFAA 2007. Lecture Notes in Computer Science, vol 4443. Springer, Berlin, Heidelberg
4. L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population. *Laboratory for Int’l Data Privacy (LIDAP-WP4),* 2000.
5. "U.S. population social media penetration 2017 | Statistic", Statista, 2017. [Online]. Available: https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/.
6. "Yelp Dataset", Yelp.com, 2018. [Online]. Available: https://www.yelp.com/dataset/documentation/json.
7. "ELKI Data Mining Framework", Elki-project.github.io. [Online]. Available: https://elki-project.github.io/.
8. A. Nawroth, "Social Networks in the Database: Using a Graph Database", Neo4j Graph Database Platform, 2009. [Online]. Available: https://neo4j.com/blog/social-networks-in-the-database-using-a-graph-database/.

Questions?