Towards Detecting Social Circles on Twitter

Arda Ünal

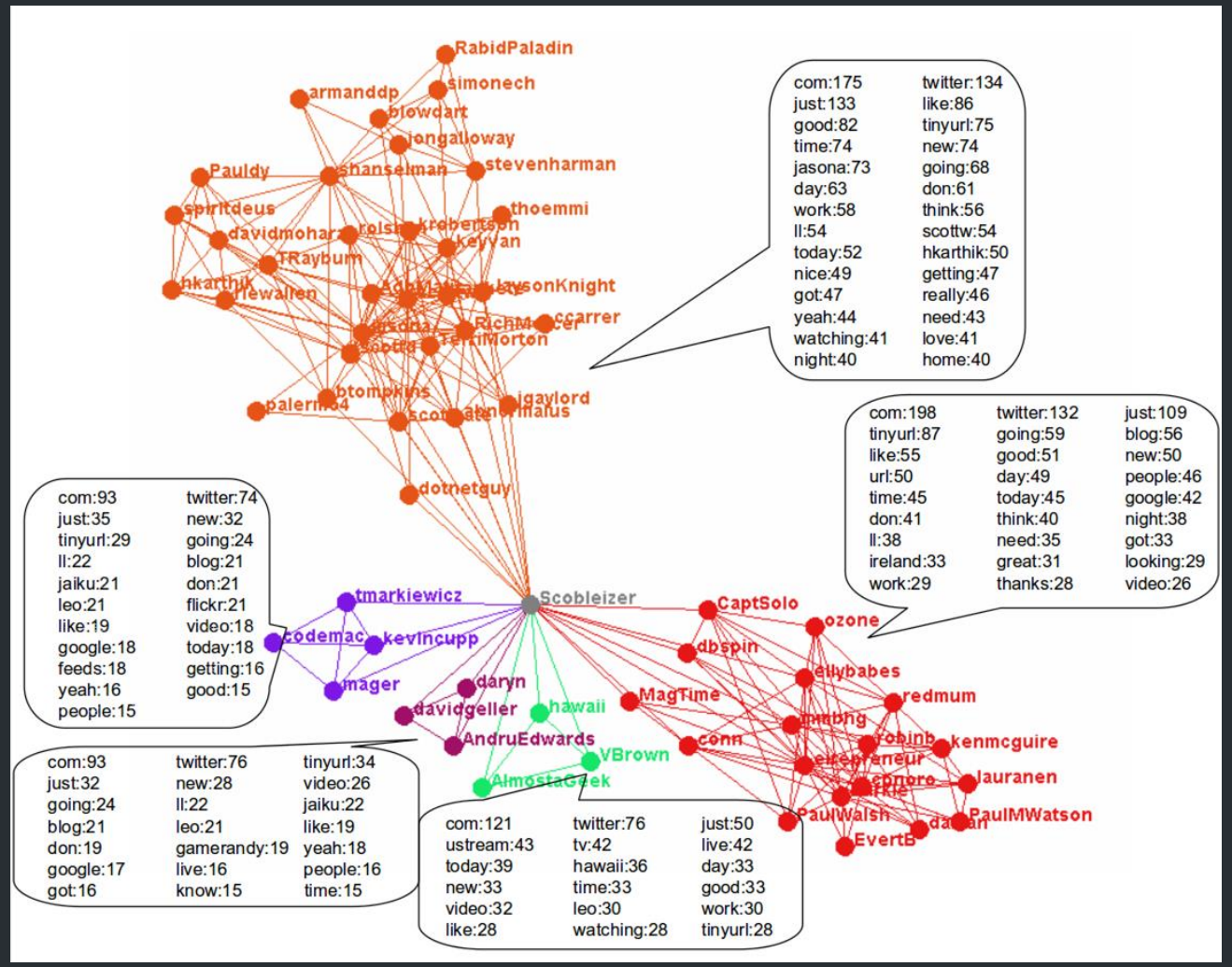
Hüseyin Celal Öner

Outline

- Problem Definition
- Proposed Methodology
- Data Set
- Expected Results

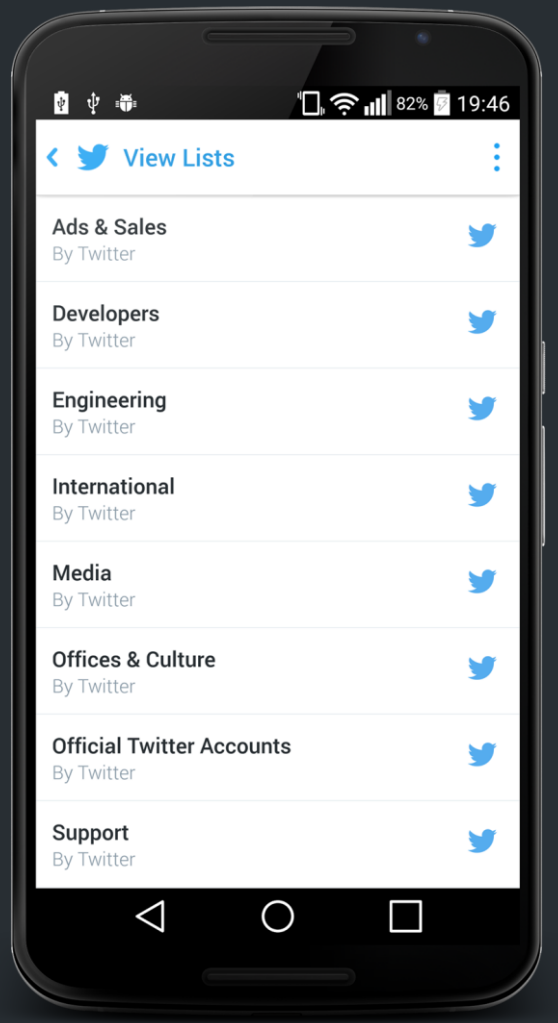
How Do People Use Twitter?

- Many Twitter users follow friends, family, celebrities, sport teams or even brands.



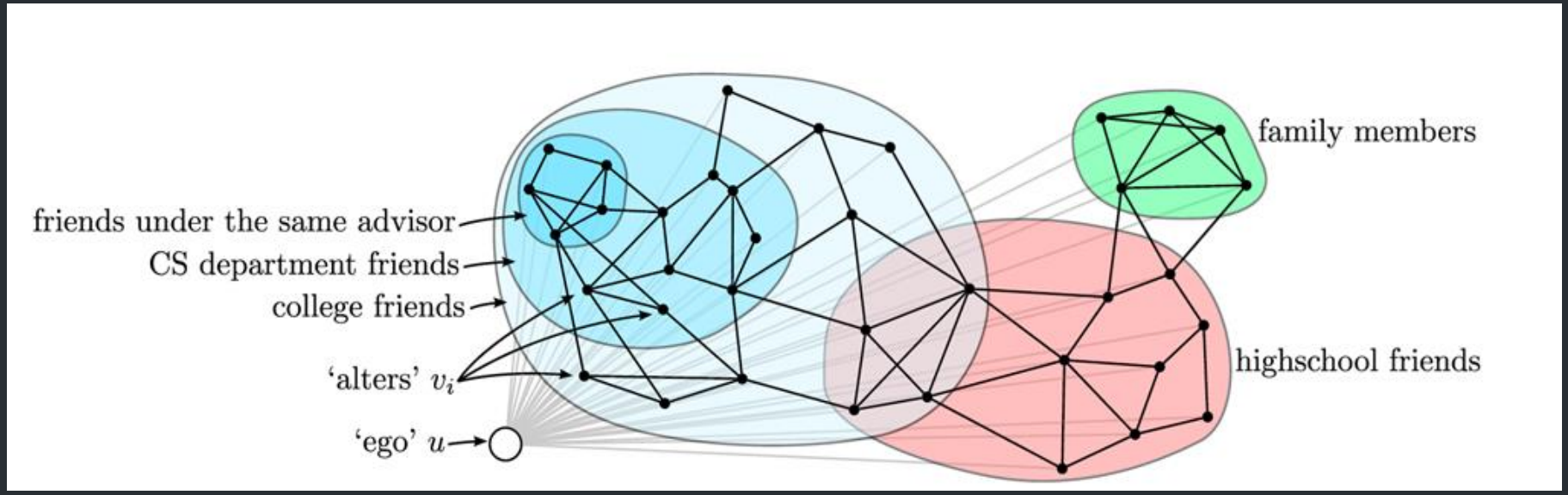
Problem Definition

- The personal social network of many Twitter users is big and untidy.
- In 2009, Twitter released a feature, “lists”
  - Designed to provide a better way to organize information on Twitter.
  - Using lists to organize information on Twitter is tedious task



Project Goal

- Automatically infer users' social circles using the set of connections between users and their social network profiles.



Proposed Methodology

1. Smooth out inconsistencies in the basic adjacency matrix so that it gives a better notion of the degree of connectedness of various node. (Exponential Adjacency Matrix)
2. Apply clustering. (K-means and Spectral Clustering)
3. Eliminate some circles found by the clustering algorithm if their density is smaller than a threshold
4. Add more people to the circles if they have number of friends over a threshold given in that circle
5. Merge circles if they overlap more than a given threshold

Data Set

- High quality hand-labeled data from major social networking sites (Facebook, Google+, Twitter)
- All data are available on snap.stanford.edu/data/

	# of ego-networks	# of social circles	# of nodes	# of edges
Facebook	10	193	4,039	88,234
Google+	133	479	107,614	13,673,453
Twitter	1000	4,869	81,306	1,768,149

Data Set

- Facebook data are fully labelled
- Google+ and Twitter data are only partially labelled
  - Collected publicly available circles(Google+) and lists(Twitter)

Features

- From Google+ they collect data from six categories (gender, last name, job titles, institutions, universities, and places lived).
- From Facebook they collect data from 26 categories, including hometowns, birthdays, colleagues, political affiliations, etc.
- For Twitter, many choices exist as proxies for user profiles; they simply collect data from two categories, namely the set of hashtags and mentions used by each user during two-weeks' worth of tweets.

Expected Results

- Lack of explicitly available features on Twitter
- Twitter data set is partially-labeled
- Expecting our methodology to work better for Facebook

References

- A. Java, X. Song, T. Finin and B. Tseng, "Why we twitter", Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07, 2007.
- "Soon to Launch: Lists | Twitter Blogs", Blog.twitter.com, 2009. [Online]. Available: https://blog.twitter.com/2009/soon-to-launch-lists.

Thank You