# Interactions of the Twitter 'Elite':
## Clustering Celebrities, Media Outlets, and Politicians

Cathy Chen
Advisor: Andrea LaPaugh

## Motivation

- Social media increasingly influences real-world events.
- Some suggest that distinctions between celebrities, media outlets, and politicians have become blurred.
- Twitter provides publicly available interactions between hundreds of millions of active users.
  - We can study how these three groups relate to each other and how distinct they are in terms of online interactions.

## Background

- Among multiple user types, information spreads homogeneously [4].
  - On Twitter, URLs tend to travel between members of the same groups.
  - Study included celebrities, media outlets, blogs, and organizations.
- Overlap of user types' roles in the real world.
  - Politicians and celebrities' roles have become intertwined [3].
  - High-profile "opinion leaders" act as a bridge between media outlets and the general public [1].
- Community Detection within Networks [2].
  - Spectral clustering uses an affinity matrix to cluster nodes in high-dimensional, non-Euclidean spaces.

## Questions of Interest

- Can we distinguish between categories of highly-visible users based on their interactions with each other?
  - Possible categories: celebrity, media outlet, politician.
  - Define "Highly-visible" users as the users with the highest number of followers.
- Which types of users interact the most with each other?
  - How does this differ across different types of interactions?

## Acknowledgments

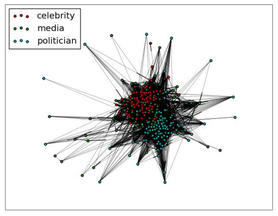
Thank you to Professor Andrea LaPaugh and to the Princeton Computer Science department.

## Approach

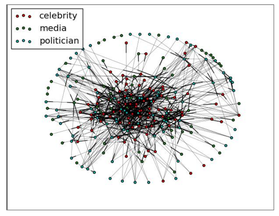
- Create interaction networks between celebrities, media outlets, and politicians with the highest number of followers (100 of each).
  - Create a separate network for each type of interaction: mentions, retweets, and replies.
- Use network metrics to measure mixing between user types.
- Cluster nodes according to interactions.
- Evaluate clustering based on strength of clusters and how well clusters capture user type labels.

## Interaction Networks

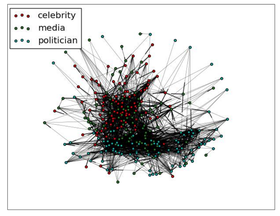
- **Mentions**: the source user links to the target user in a Tweet.
- **Replies**: the source user replies to one of the target user's Tweets.
- **Retweets**: the source user re-posts one of the target user's Tweets.




Mentions Network



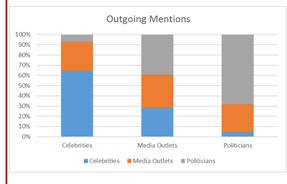

Replies Network

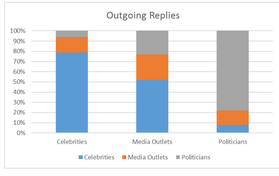


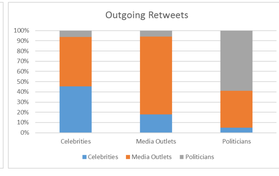

Retweets Network

## Network Metrics

- Mixing, separated by user type: mainly homogeneous for celebrities and politicians, more heterogeneous for media outlets.




- Mixing for each type of network.
  - Attribute assortativity coefficient measures amount of mixing between different types of users. Value of 1 indicates perfect homogeneity, 0 indicates random distribution of interactions between user types.

| Interaction Type | Attribute Assortativity Coefficient |
|---|---|
| Mentions | 0.28 |
| Replies | 0.55 |
| Retweets | 0.32 |

## Clustering

- We use $A + A^{\top}$ symmetrization of the network adjacency matrix as the affinity matrix used for Spectral Clustering.
  - Network adjacency matrix: $A_{\{ij\}} = 1$ if user *i* initiates an interaction with user *j*.
  - Chosen based on lowest conductance score after testing various symmetrization methods (bibliometric, degree-discounted, undirected adjacency matrix).
- Spectral Clustering projects affinity matrix to lower-rank matrix and performs k-means clustering on projection.

## Cluster Evaluation

- Evaluation Metrics
  - Conductance: strength of separation between clusters (0 indicates no crossing edges between clusters, 1 indicates that all edges from nodes in one cluster go to nodes in another cluster).
  - Homogeneity: how well clusters separate members of different categories.
  - Completeness: how well clusters group members of the same category.
  - V-measure: metric that balances homogeneity and completeness.
- For comparison, also evaluated clusters produced using adjacency matrix of undirected interaction graph as the affinity matrix for spectral clustering.

| | $A + A^{\top}$ | | | Adjacency Matrix of Undirected Graph | | |
|---|---|---|---|---|---|---|
| | Mentions | Replies | Retweets | Mentions | Replies | Retweets |
| Conductance | 0.295 | 0.153 | 0.114 | 0.658 | 0.608 | 0.662 |
| Homogeneity | 0.381 | 0.356 | 0.015 | 0.004 | 0.015 | 0.035 |
| Completeness | 0.398 | 0.410 | 0.017 | 0.004 | 0.017 | 0.035 |
| V-measure | 0.389 | 0.015 | 0.016 | 0.004 | 0.016 | 0.035 |

## References

1. S. Choi, "The two-step flow of communication in twitter-based public forums," Social Science Computer Review, 2014.
2. M. Porter, J.-P. Onnela, and P. Mucha, "Communities in networks," Notices of the American Mathematical Society, vol. 56, no. 9, pp. 1082–1097, 1164–1166, 2009.
3. J. Street, "Celebrity politicians: Popular culture and political representation," The British Journal of Politics and International Relations, vol. 6, no. 4, pp. 435–452, 2004.
4. S. Wu et al., "Who says what to whom on twitter," in Proceedings of the 20th International Conference on World Wide Web, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 705–714.