A. Debruyn, S. Gettys
Faculty Advisors: Dr. Feng Wang
School of Mathematical and Natural Sciences

Community Detection and Verification in Large Scale Social Media

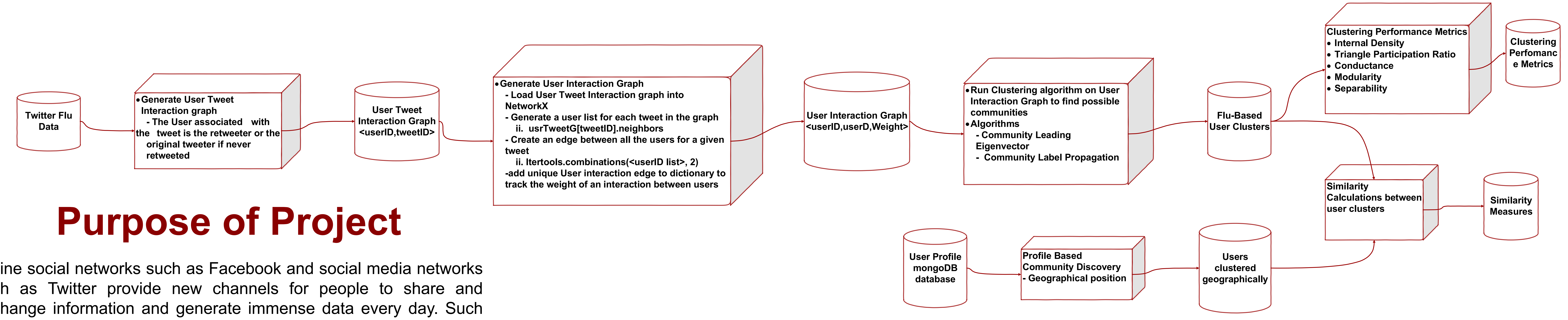
Twitter Flu Data
• Generate User Tweet Interaction graph
- The User associated with the tweet is the retweeter or the original tweeter if never retweeted
User Tweet Interaction Graph <userID,tweetID>
• Generate User Interaction Graph
- Load User Tweet Interaction graph into NetworkX
- Generate a user list for each tweet in the graph
ii. usrTweetG[tweetID].neighbors
- Create an edge between all the users for a given tweet
ii. Itertools.combinations(<userID list>, 2)
-add unique User interaction edge to dictionary to track the weight of an interaction between users
User Interaction Graph <userID,userD,Weight>
• Run Clustering algorithm on User Interaction Graph to find possible communities
• Algorithms
- Community Leading Eigenvector
- Community Label Propagation
Flu-Based User Clusters
Clustering Performance Metrics
• Internal Density
• Triangle Participation Ratio
• Conductance
• Modularity
• Separability
Clustering Perfomance Metrics
Similarity Calculations between user clusters
Similarity Measures
User Profile mongoDB database
Profile Based Community Discovery
- Geographical position
Users clustered geographically

Figure 1: Twitter Flu Data Clustering Flow Diagram

Purpose of Project

Online social networks such as Facebook and social media networks such as Twitter provide new channels for people to share and exchange information and generate immense data every day. Such networks often exhibit community structure with inherent clusters. Detecting clusters or communities is one of the critical tasks in social network analysis due to its broad applications such as friend recommendations, link predictions and collaborative filtering. We want to study the community detection and verification problem through user tweet interaction and since we have built a data collection, analysis, and modeling frame work, we have complete control of our data and can collect any additional data, so we are able to verify our clustering approach with real data.

Framework

As illustrated in Figure 1, the raw data consists of Tweets posted by users of Twitter regarding the flu season of 2014 (Jan – Mar). We then strip down the raw tweets to the tweet ID number and the User ID number of the person that retweeted the tweet, or the original tweeter if the prior is not present, using the Dynamic Modeling Engine Jaime Chon developed. From there we develop a graph of users and tweets with connections between them called edges. The retrieving the neighbors of all the tweet nodes we can get users that have tweeted or retweeted the tweet and construct an edge list between users, which will be the raw data for our clustering algorithms in the iGraph library. We then will test our clusters against clusters generated from user profile data to measure the accuracy of our flu data based clusters.

Conclusion

The clusters that we produce and analyze add a new angle of cluster analysis that is not being utilized in the social media industry, clustering based on user behavior. Being able to analyze user behavior at this level, allows us to predict trends and possible undiscovered clusters that are not being caught with current industry standards of analysis.

Clustering

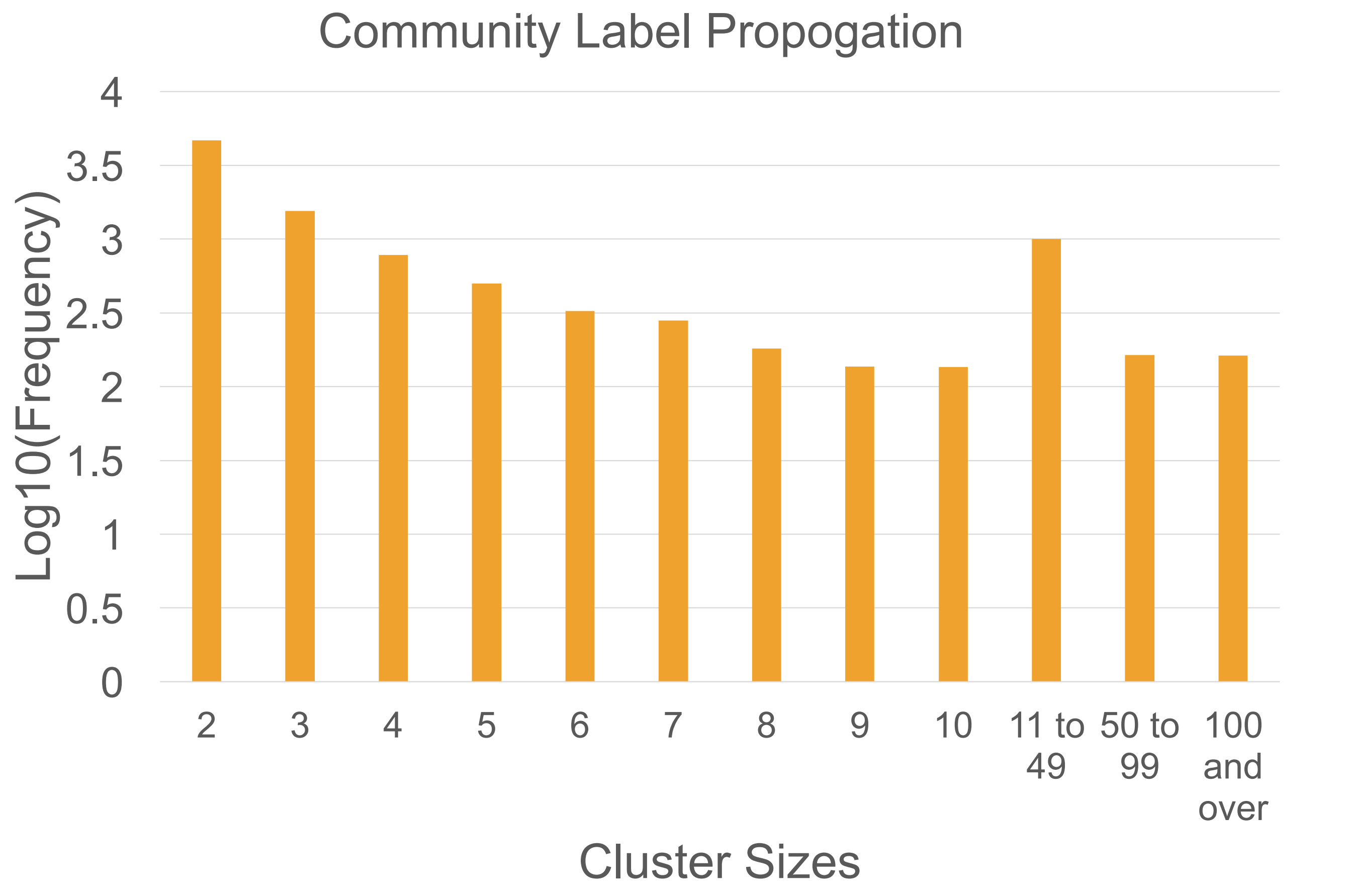

Figure 4: Community Label Propagation clustering results with weighted edges showing the number of clusters of different sizes

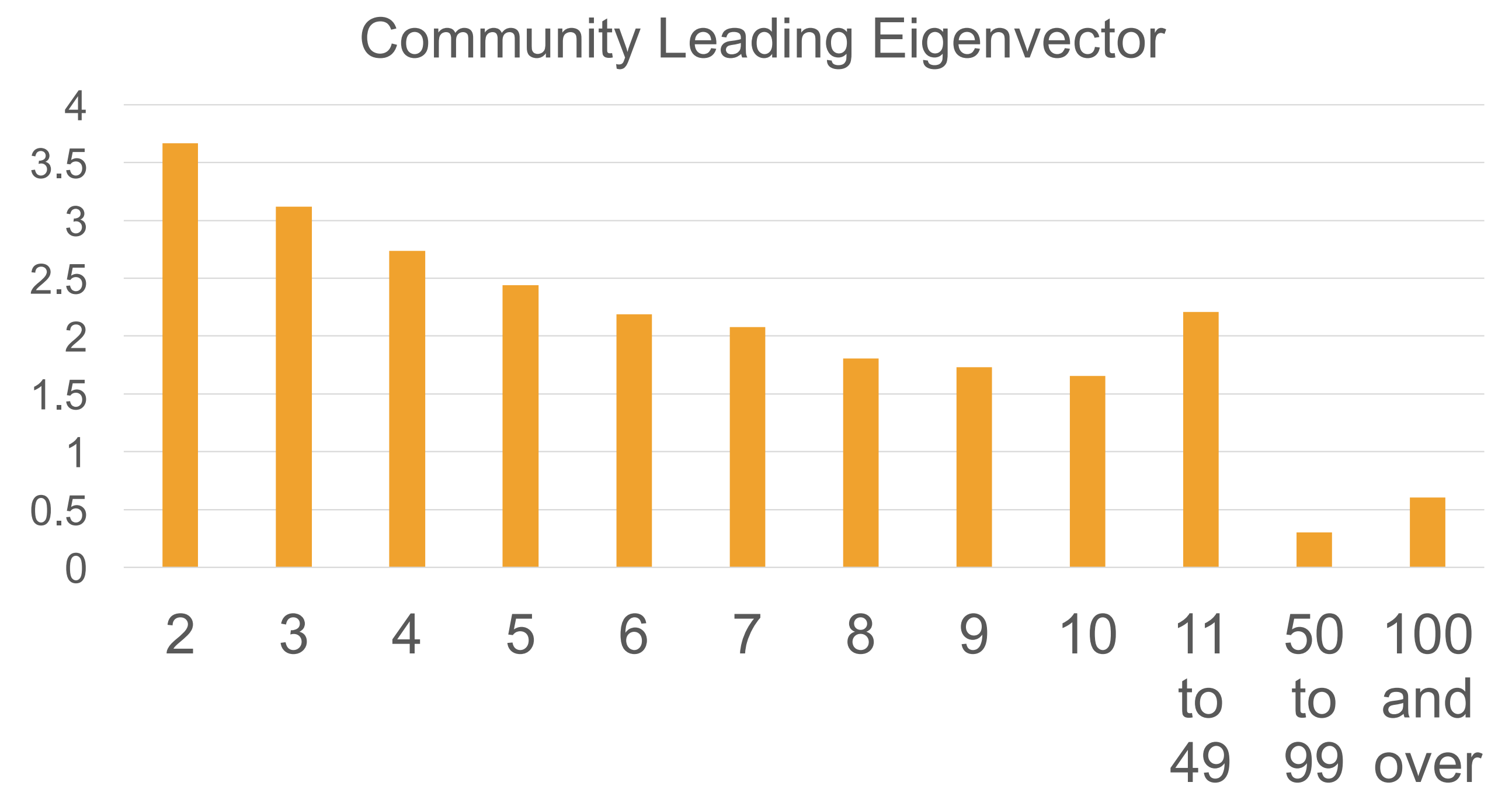

Figure 3: Community Leading Eigenvector clustering results with weighted edges showing the number of clusters of different sizes

User Interaction Weight Analysis

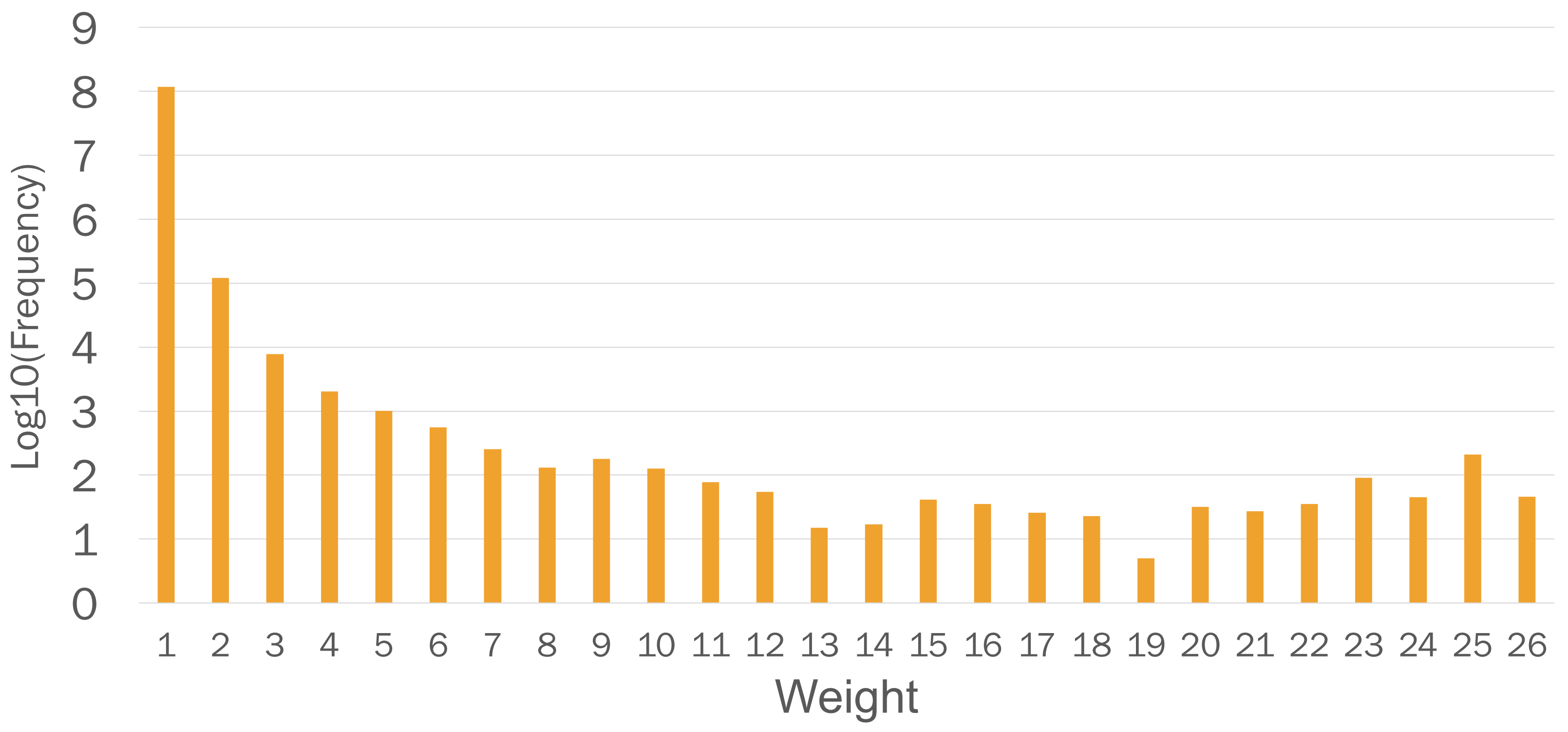

Figure 2: The weight of a user interaction depicts the number of tweets a pair of users have in common

References

- Modeling Flu Trends with Real-Time Geo-tagged Twitter Data Streams, Jaime Chon, Ross Raymond, Haiyan Wang, and Feng Wang, Wireless Algorithms, Systems, and Applications, 2015
- The Impact of Sampling on Big Data Analysis of Social Media: A Case Study on Flu and Ebola, Kuai Xu, Feng Wang, Xiaohua Jia, Haiyan Wang, IEEE GlobeCom 2015
- Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 76(3), 036106.
- Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. Physical review E, 74(3), 036104.

Acknowledgements

- "Dynamic Mathematical Modeling Towards Understanding Information Diffusion in Online Social Networks", NSF
- "Minimal Cost and Time-sensitive Information Diffusion in Online Social Networks", HK RGC GRF