ASU new college
ARIZONA STATE UNIVERSITY

P. Wagenseller III
Faculty Advisors: Dr. Feng Wang
School of Mathematical and Natural Sciences

# Size Matters: Empirical Evaluation of Community Detection Algorithms

## Introduction

Understanding community structures in social media based networks we can use the information for link predictions, social bot detection, friend recommendations, collaborative filtering, and more. While no formal definition of a community exists, communities can be thought of as groups of nodes within a graph or network that have more connections amongst themselves then to the surrounding graph. The algorithms designed to detect these structures often attempt to maximize or minimize a certain metric such as modularity.

In this project, we compared the widely-known algorithms of Infomap, Multilevel, Eigenvector, and Fastgreedy against a baseline algorithm we designed called the Clique Augmentation Algorithm (CAA). We discovered that most communities found by these popular community detection algorithms do not take community size into account. Recent research [1] by Robin Dunbar suggests that community size of a community should be limited to 150 individuals due to cognitive and time constraints of humans both in online and offline social networks. While small communities of size less than 3 individuals are often trivial and not is what is typically thought of as a community.

## Clique Augmentation

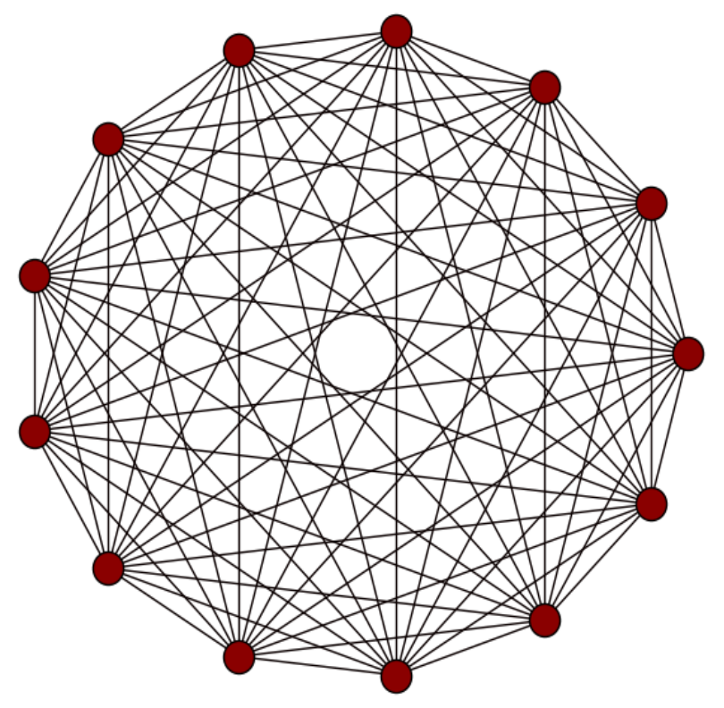

Figure 1 – Clique

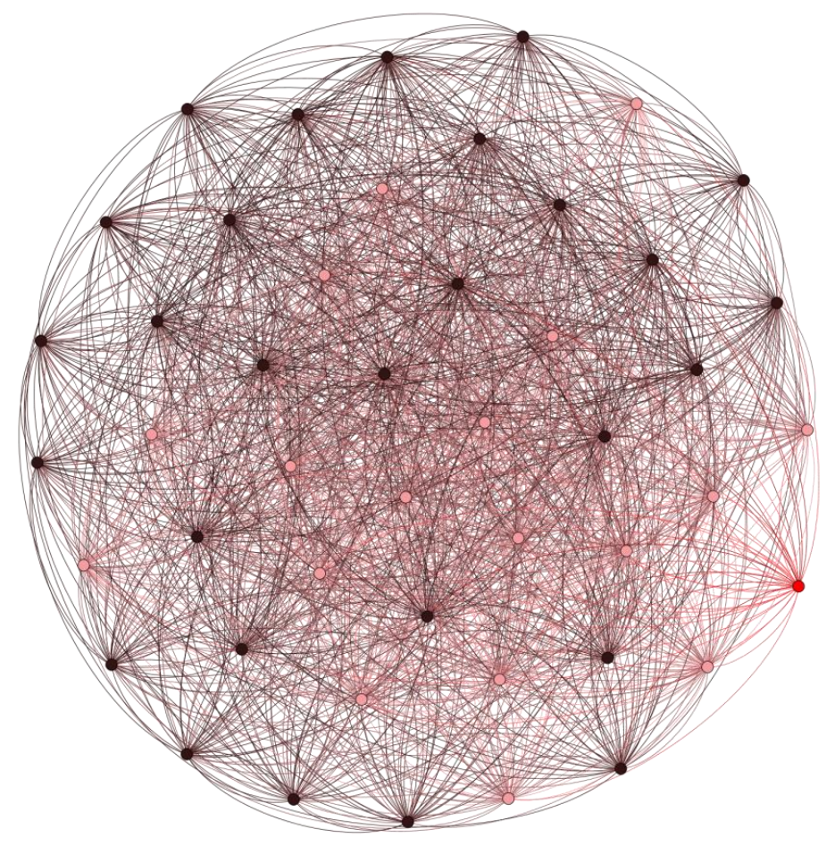

Figure 2 – Community

The clique augmentation algorithm (CAA) is the baseline algorithm we proposed to compare against other algorithms. The way it works is as follows:

1. We start by finding all cliques an example of a clique can be seen in Figure 1. This is done using the Bron-Kerbosch algorithm.

2. Filter the cliques allowing for a certain amount of overlap between nodes of the cliques controlled using what is called the overlapping threshold. The overlapping threshold is the percentage of overlapping nodes in the smaller of the two cliques.

3. Grow each clique into a community. This is done using a value called the growing threshold. The growing threshold is the ratio of edges that must be connected to the current community for a node to be added to the community. For example, a value of 0.6 indicates a node must be connected to at least 60% of the nodes in the current community's iteration.

## Size Distribution

In Figure 3 we see what percentage of communities fall within various size ranges. We can see that most communities fall within the very small 1-3 size range.

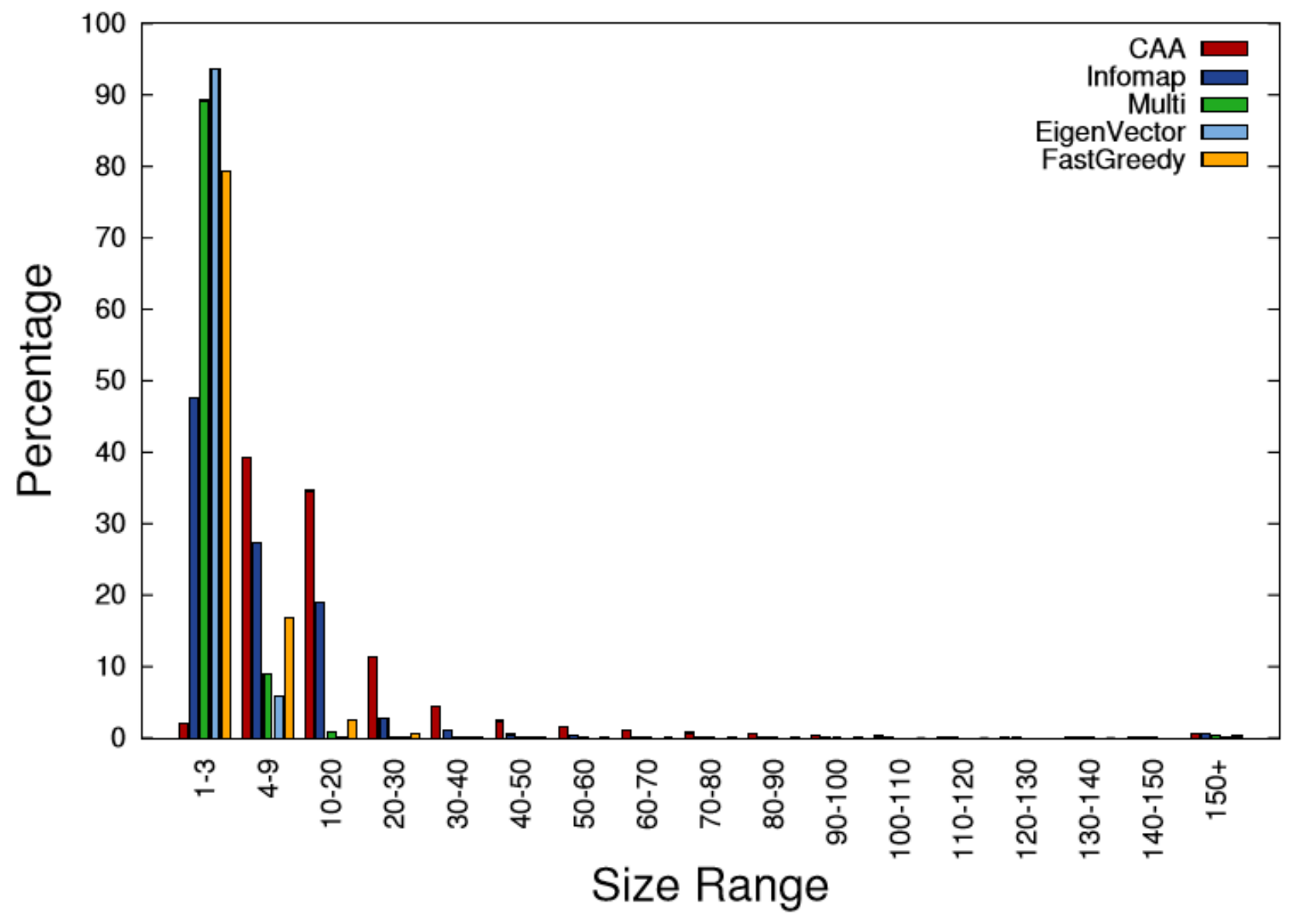

Figure 3 – Size Distribution

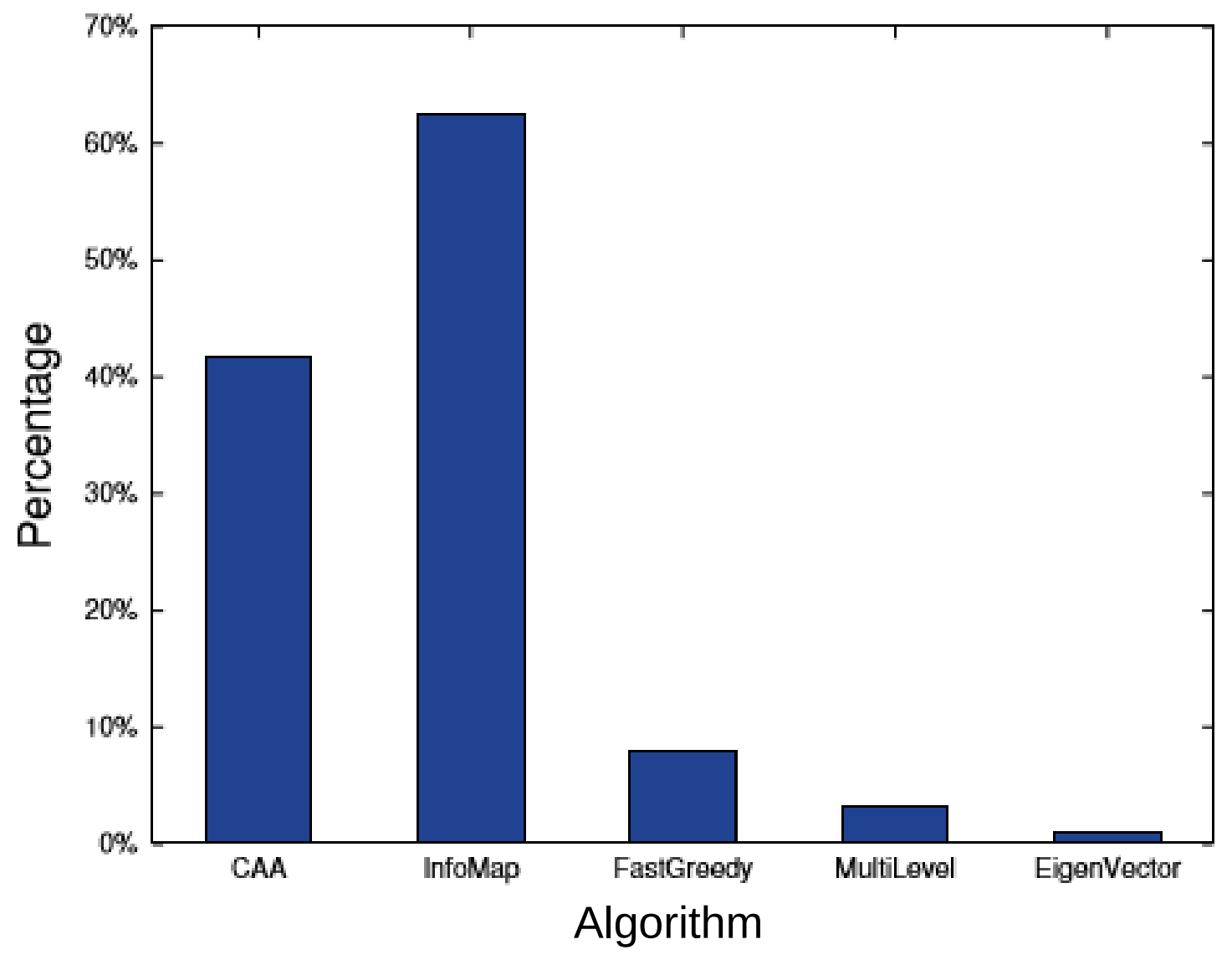

Figure 4 – Community Coverage

Many community detection algorithms claim to correctly identify the community for any node within the network providing 100% coverage. However, when you consider size of the communities discovered you find they do not provide full coverage as seen in Figure 4.

## Modularity

Modularity is a traditional metric that measures the division of a network into modules. We divide the modularity result into different size ranges to see the overall contributing amount in Figure 5. Higher modularity typically indicates better communities. The definition of modularity can be seen in Equation 1.

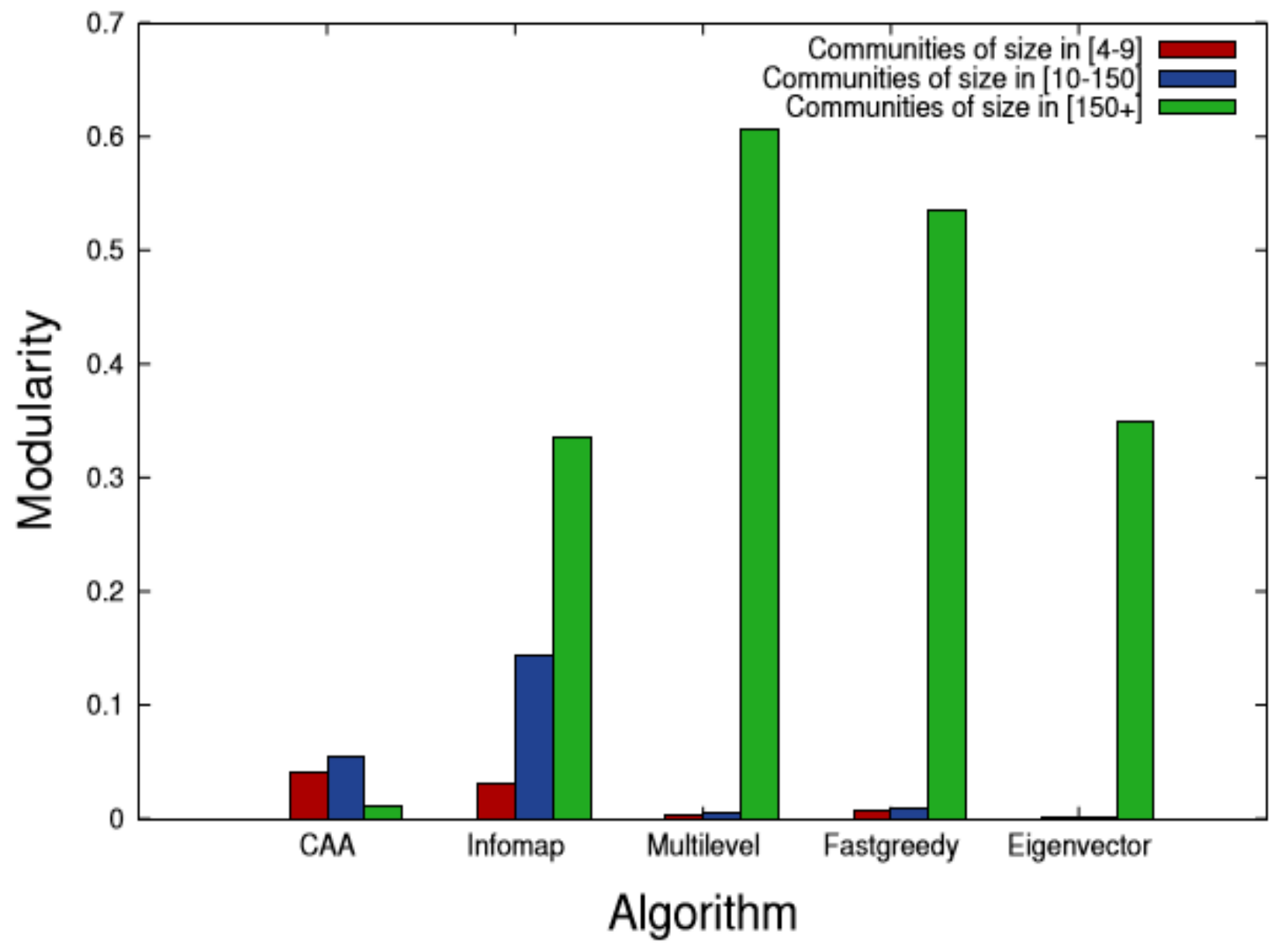

Figure 5 – Extended Modularity

$$EQ = \frac{1}{2m}\sum_{i}\sum_{v\in c_i, w\in c_i}\frac{1}{O_v O_w}\left[A_{vw} - \frac{k_v k_w}{2m}\right]$$

Equation 1 – Definition of Extended Modularity

- $O_v$is the number of communities the vertex *v* belongs to.
- $O_w$is the number of communities the vertex *w* belongs to.
- $A_{vw}$ = 1 when there is an edge between vertex *v* and *w*.
- $\frac{k_v k_w}{2m}$ is the expected number of edges between vertex *v* and *w*.
- $k_v$ is the degree of vertex v.
- $k_w$ is the degree of vertex w.
- $m$ is the total number of edges within the topology.

## Additional Metrics

We investigated many different metrics to verify the community quality of the communities found. During our research, we have evaluated the metrics of modularity, triangle participation ratio, conductance, internal density, and transitivity with relation to community size. We found that our baseline algorithm CAA, performed best in most of these metrics.

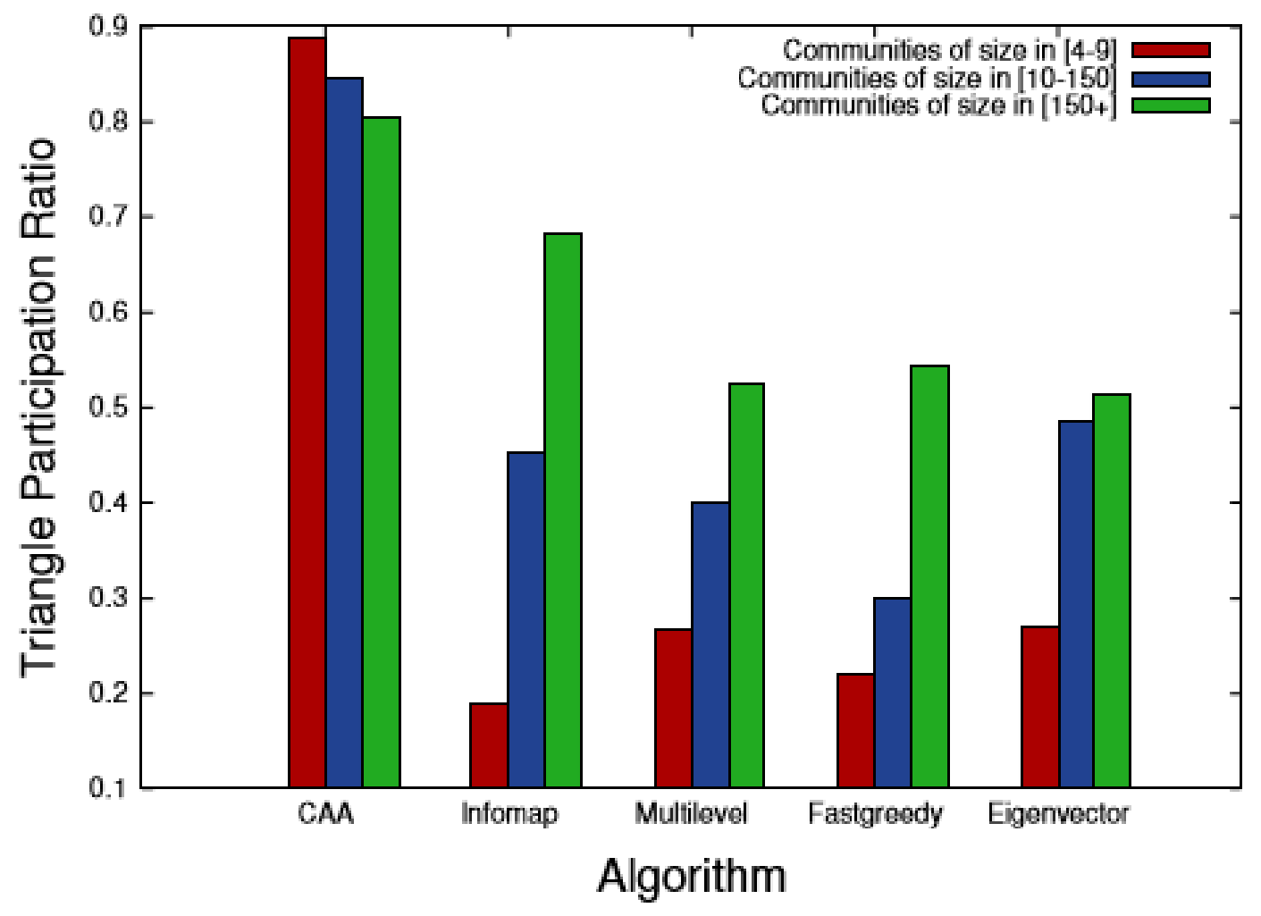

Figure 6 – Triangle Participation Ratio

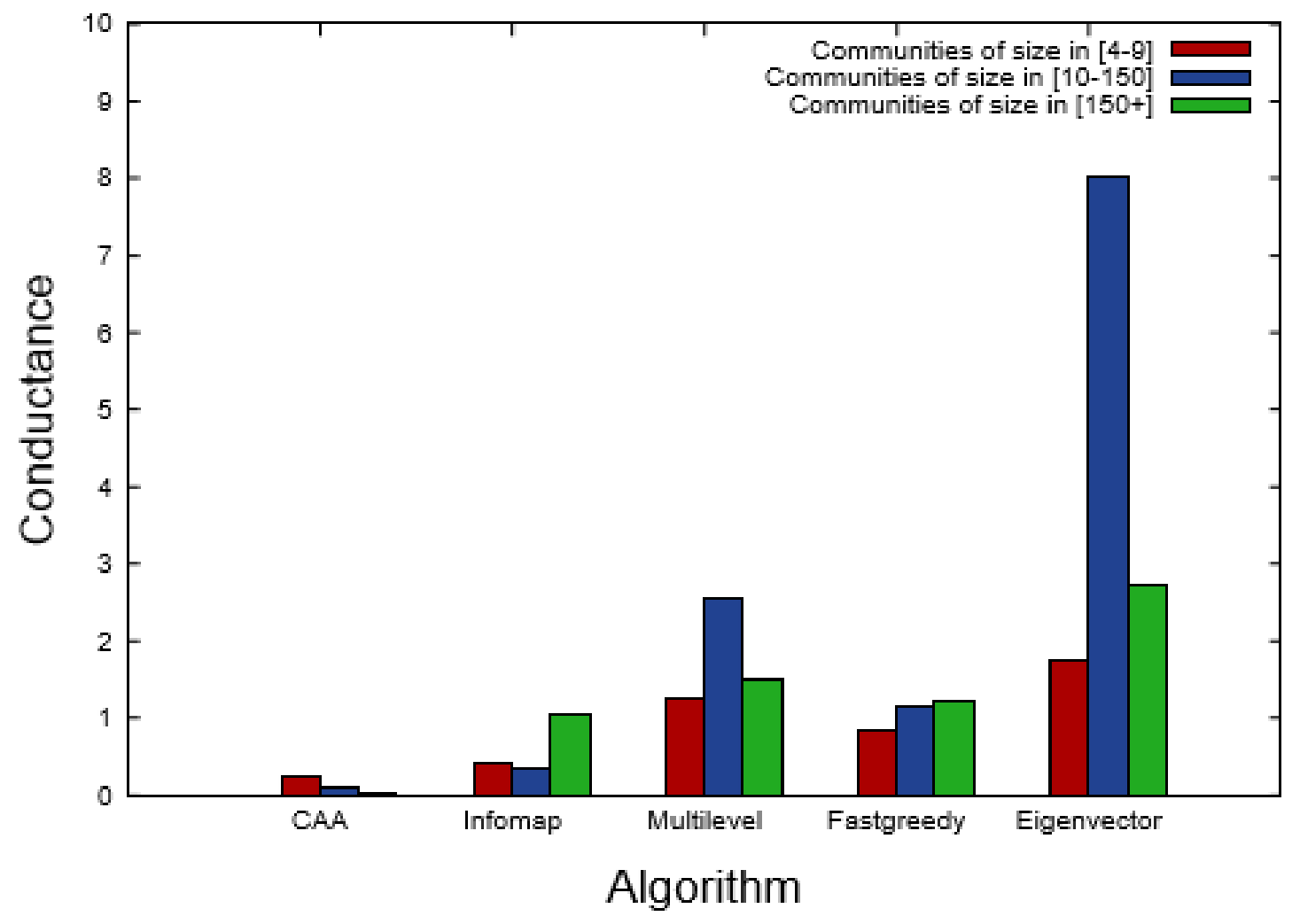

Figure 7 – Conductance

As seen in Figure 6 the triangle participation ratio for the baseline algorithm CAA is extremely high. The triangle participation ratio is defined as the number of nodes in a community that form a triad divided by the total number of nodes in the community. The reason CAA performs so well is because our communities are grown from cliques. Where a clique is defined as a complete graph such that every node is connected to every other node.

Conductance is defined $\frac{c_S}{2m_s+c_S}$ where $c_S$ is the number of edges in community $S$ and $2m_s + c_S$ is the number of edges leaving the community. Lower conductance score indicates a better community. We found that the conductance value for the baseline algorithm was extremely low as seen in Figure 7.

## Conclusion

We studied community detection algorithms and metrics involving the quality of their result with respect to community size. Our results indicate that Infomap and our own CAA algorithm can discover meaningful communities. Meanwhile existing algorithms often find too large or too small communities.

In the future, we plan on using these findings to improve community detection algorithms. Analyze the content posted by the users in Twitter and compare the result to the structure of the graph. Use community detection algorithms to address the social bot and fake news problem.

## References

[1] Dunbar, R. I. M. "Do Online Social Media Cut through the Constraints That Limit the Size of Offline Social Networks? " in *R. Soc. Open Sci. Royal Society Open Science 3.1 (2016): 150292.*

[2] Paul Wagenseller III, and Feng Wang. "Community Detection Algorithm Evaluation Using Size and Hashtags." *[1612.03362] Community Detection Algorithm Evaluation Using Size and Hashtags., 11 Dec. 2016. Web. 12 Apr. 2017. <https://arxiv.org/abs/1612.03362>.*