Ken Orton Feng Wang
School of Mathematical and Natural Sciences
New College of Interdisciplinary Arts and Sciences
Arizona State University

Community Verification using Topic Modeling

Introduction

Users on social media platforms like Twitter create reciprocal friendships with each other, often forming groups of interconnected follower relationships. In the context of network theory, these relationships between groups of connected users are referred to as communities. There exist many different algorithms for discovering communities but little research has been done to understand the significance of the context of association within these communities. A simple approach to discovering the context of user relationships would be to read through the tweets of each user in the community and come to a conclusion about the interests of each user based on the observations. Unfortunately, this approach will not scale well for large amounts of data and the categorization of interests would be biased in regards to the person who is interpreting the information. In this project, we examine whether or not users in communities discovered by a topology-based community detection algorithm display similar interests. We use a topology of communities already discovered using an algorithm called Clique Augmentation Algorithm and we generate an LDA model using Wikipedia as a training set to identify latent topics in the tweets of each user in the communities.

Topic Model

Latent Dirichlet Allocation, or LDA, is a generative statistical modeling approach where topics are derived from a corpus of known training data, which provides a mechanism for predicting the distribution of topics of unseen documents.

LDA
num_topics = 2

Corpus	
Document	Words
1	apple, banana
2	apple, orange
3	banana, orange
4	tiger, cat
5	tiger, dog
6	cat, dog

Documents of Interest	
Document	Words
7	cat, dog, apple

Word topic distribution		
word	topic 1	topic 2
apple	33%	0%
banana	33%	0%
orange	33%	0%
tiger	0%	33%
cat	0%	33%
dog	0%	33%

Document topic distribution		
Document	topic 1	topic 2
1	100%	0%
2	100%	0%
3	100%	0%
4	0%	100%
5	0%	100%
6	0%	100%
7	33%	66%

Figure 1: An illustration of the LDA approach courtesy of Yuhao Yang

Overview

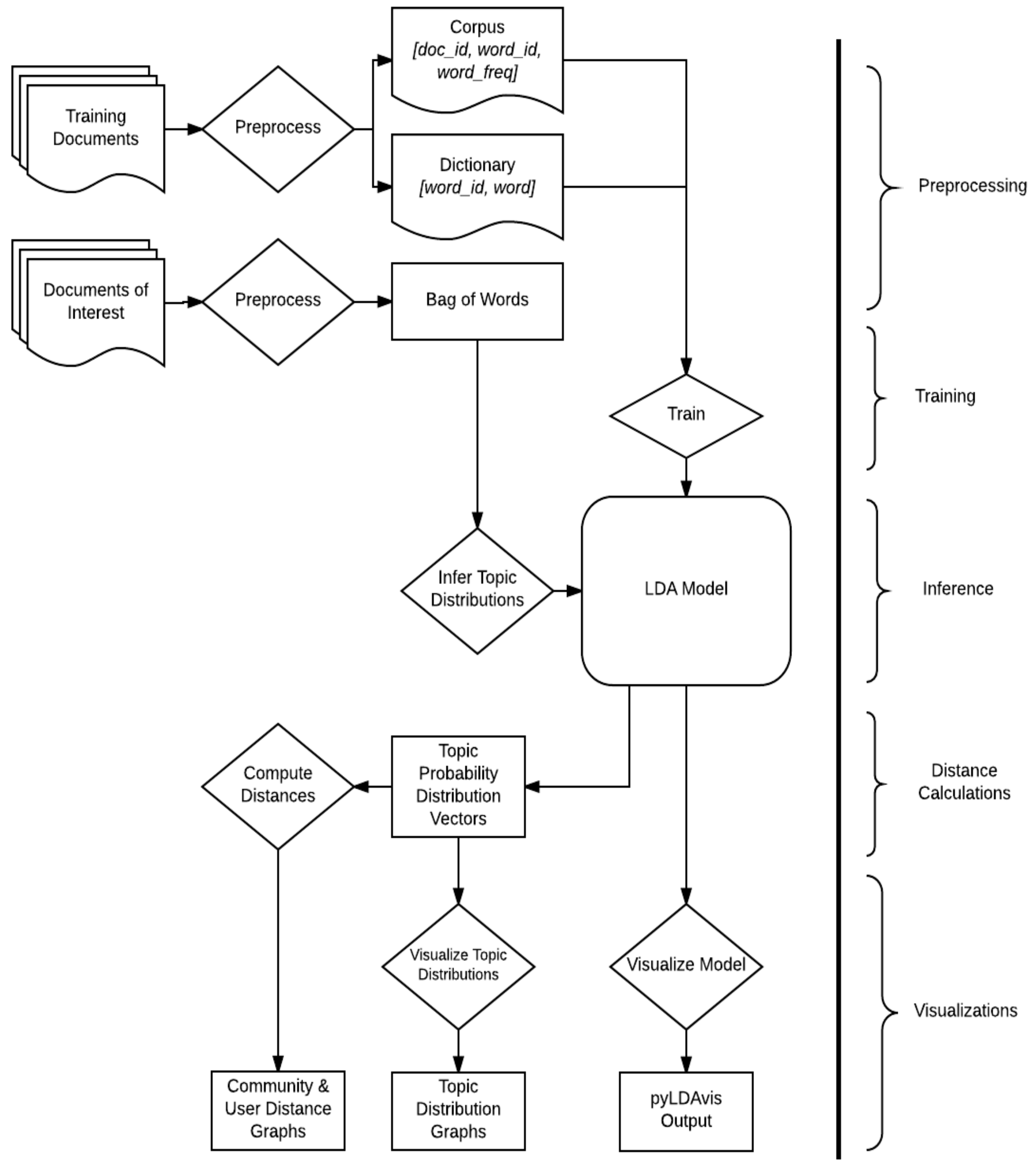
Training Documents
Preprocess
Corpus [doc_id, word_id, word_freq]
Dictionary [word_id, word]
Documents of Interest
Preprocess
Bag of Words
Train
Infer Topic Distributions
LDA Model
Compute Distances
Topic Probability Distribution Vectors
Visualize Topic Distributions
Visualize Model
Community & User Distance Graphs
Topic Distribution Graphs
pyLDAvis Output
Preprocessing
Training
Inference
Distance Calculations
Visualizations

Figure 2. Topic Modeling Architecture

Preprocess Training Documents: Wikipedia articles containing less than 200 words are omitted from the corpus. The words in each document are filtered using a stopword list and the words are lemmatized. The resulting dictionary is limited to a vocabulary of 170,000 words.

Preprocess Documents of Interest: All of a single users tweets make up one document. The words in each document are filtered using a stopword list. URL's and special characters are removed, and the words are lemmatized.

Training: The number of topics selected for training is 50 and the model is trained using an asymmetric prior with 5 full passes over the corpus.

Inference: Each preprocessed document of interest is used to query the trained model which generates a topic probability distribution vector for the document.

Calculations: Using the community topology as a map, each user's probability distribution vector is used to calculate similarities between other users. The similarity metric used is the Jensen Shannon divergence.

Visualizations: Each users' topic probability distribution vector is plotted and a pyLDAvis output showing the topics in the model with their top-n words is generated.

Results

A randomly selected user from the topology is investigated and established to be a bankruptcy law firm in Arizona with the username @bankruptcyazlaw. The topic probability distribution vector for the user is found and visualized in figure 3.

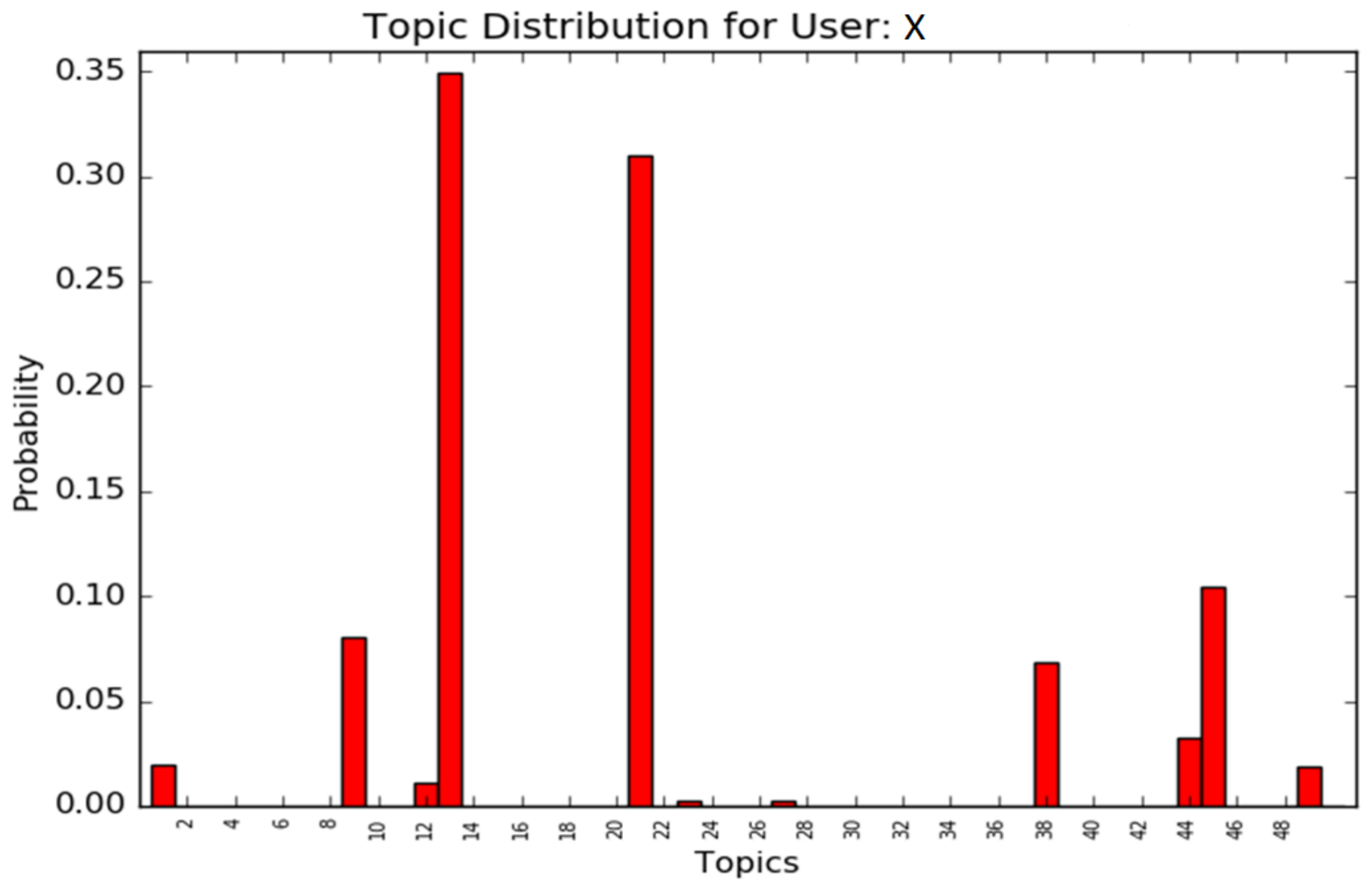

Figure 3. Topic Probability distribution for a user

The LDA model is visualized in figure 4 where each circle on the map is a topic in the model. The top 30 most relevant words in each topic are listed next to the map. The red bar indicates the frequency of occurrence of the word in the topic and the blue bar represents the frequency of occurrence of the word in the entire corpus. The distance between the topics on the map is defined by the Jensen Shannon Divergence and represents how closely related the topics are to each other in the model.

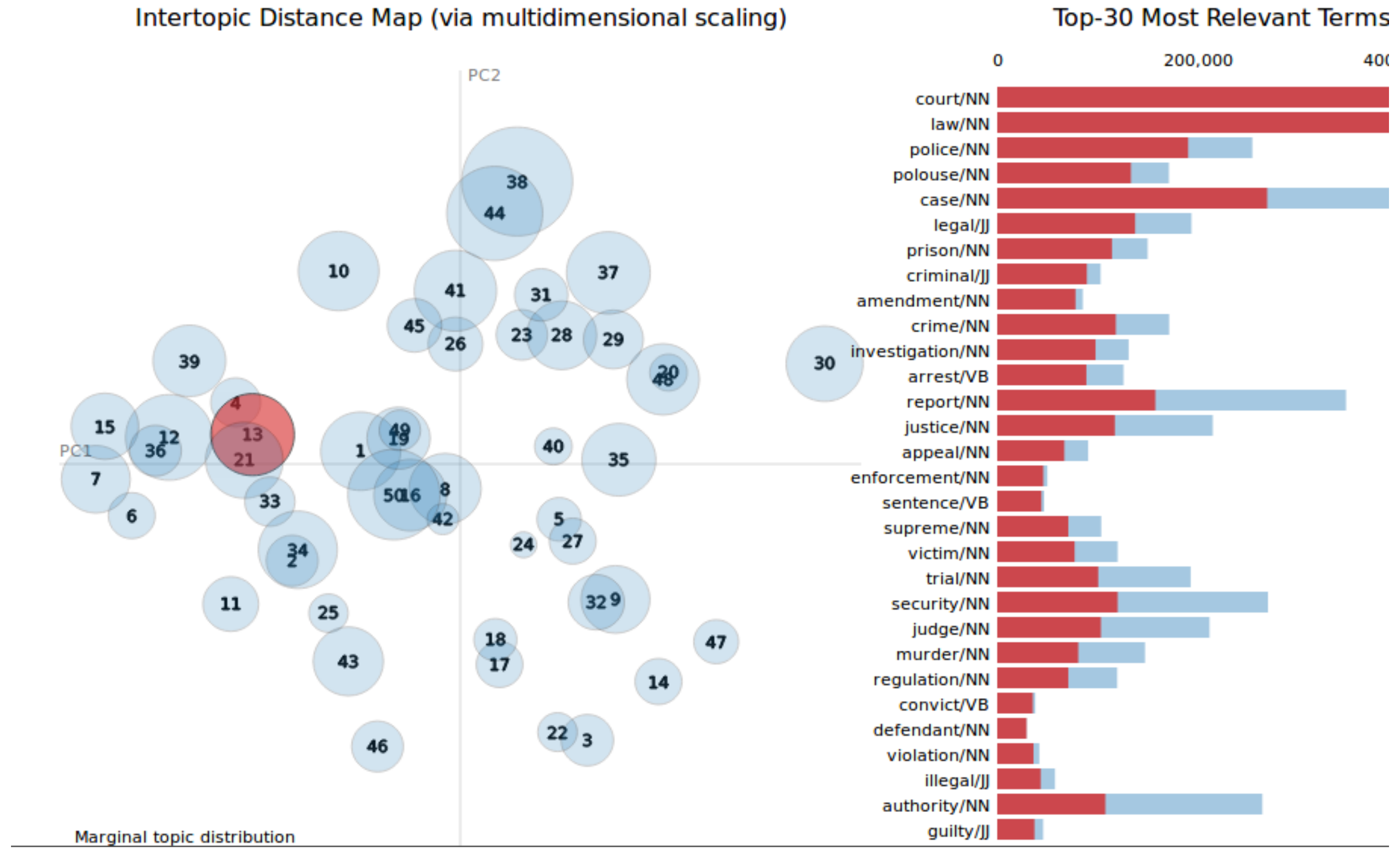

Figure 4. Visualization of topics in LDA model

The topic probability distribution vector for user in figure 3 shows topics 13 and 21 to be the most prominent. From the LDA model visualization, topic 13 contains many words about the legal system. Although not shown, topic 21 in the model contains words about economic institutions. The two topics in the LDA model visualization are also very close together on the map, indicating that they are semantically related to each other.

Reference

Feng Wang , Ken Orton "Community Verification with Topic Modeling", 12th International Conference on Wireless Algorithms, Systems, and Applications, June, 2017

Results

A community of users can be analyzed using their topic probability distribution vectors from the LDA model by calculating the distances between them. Figure 5 shows a randomly selected users shows stronger interest with users in the same community compared to users external to the community. Figure 6 shows the comparison at the community level. It can be seen that internal Jensen Shannon Divergence is clearly separated from the external JSD.

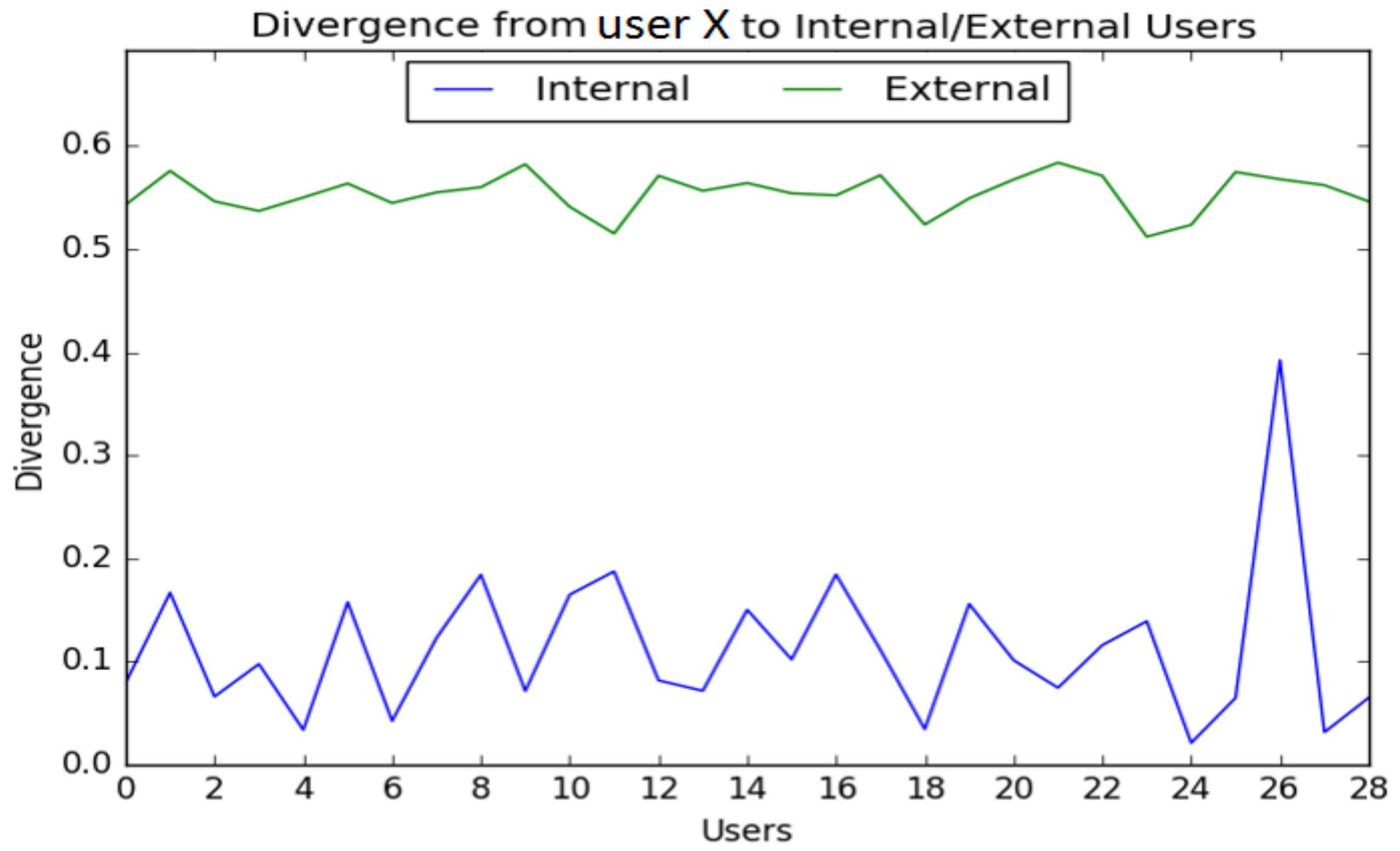

Figure 5. Jensen Shannon Divergence between a user and members of their own community vs members outside their community

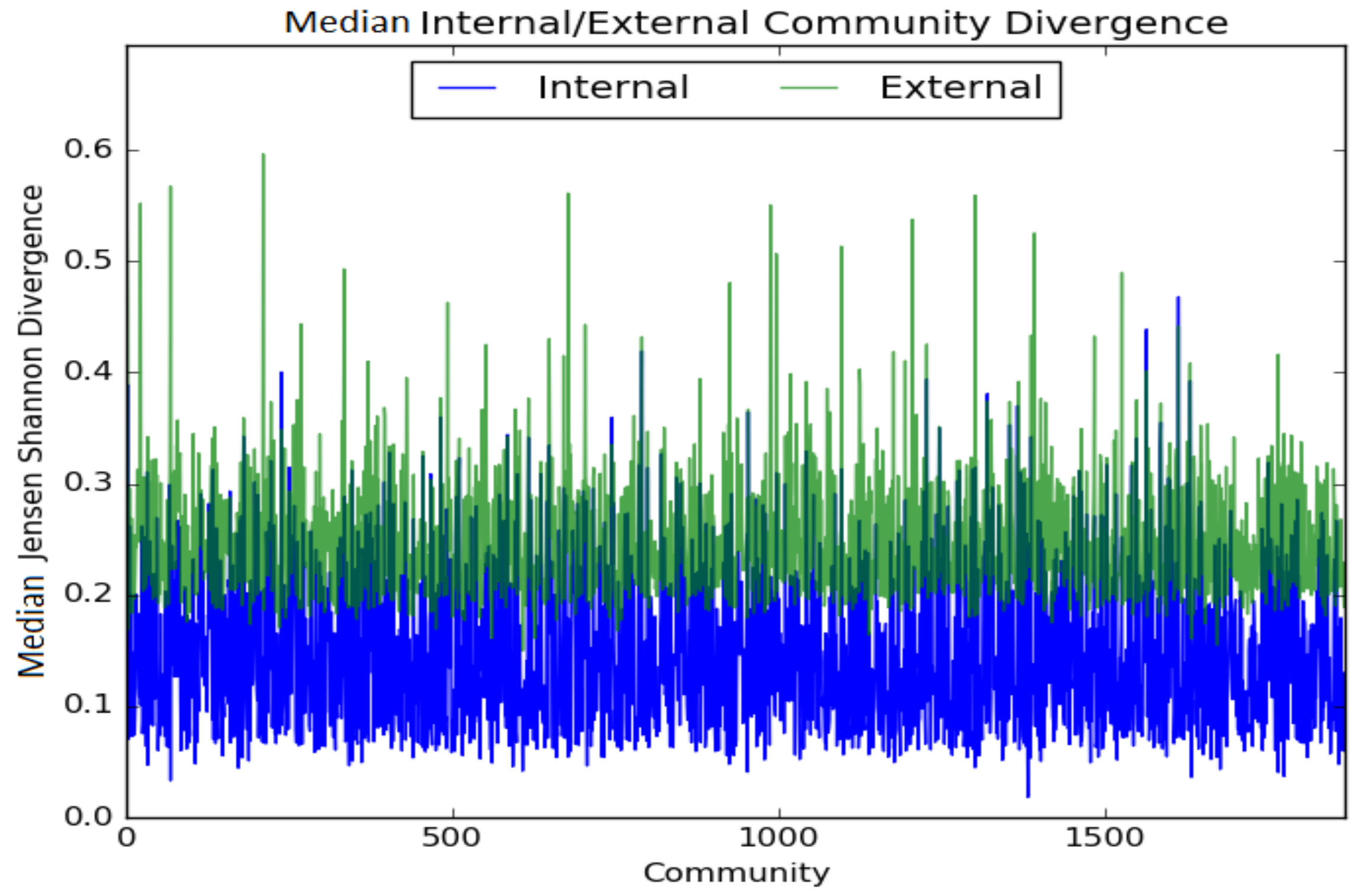

Figure 6. Average Jensen Shannon Divergence between internal users vs. external users.

Conclusion

- We propose a new methodology to validate communities based on shared interest
- We propose new measurements for community verification
  - internal and external similarity at user, clique, community level
- Clique based algorithm can detect hidden communities that show strong community theme
- We give evidence to support the statement that users form community because they share interest