Detecting Communities on Twitter Using Hashtag Networks

Daniel Mortenson

May 2022

1 Introduction

Twitter is a very vibrant, dynamic, and active forum for anyone anywhere to publish 140-character-long "tweets" on the internet for everyone to see. Twitter's user base spans almost all nations, most industries, and many ages and groups of people, and has even become the de-facto "instant news" source for governments, corporations, and celebrities.

Twitter's structure is relatively simple compared to other social media sites such as Facebook or Instagram. On Twitter, users are able to scroll through the tweets of people they follow as well as receive recommendations from Twitter's algorithm and sponsored messages. One of the most important organizational structures within Twitter is the use of hashtags. Hashtags are strings of text, starting with the hash (#) symbol, with no spaces that allow users to self-categorize their tweets with other tweets pertaining to the same subject matter that also contain the same hashtags. Searching twitter by hashtag is the easiest way to find many perspectives on the same issue, because even in contentious exchanges, users on both sides of an issue generally use the same hashtags.

The use of hashtags provides social data scientists with the ability to classify tweets without using complicated natural language processing (NLP) techniques. In this study, we analyze what sort of communities form around the use of hashtags on twitter, and whether hashtag communities contribute to curated "echo chambers" that conform to users' preferences and ideologies.

2 Data

2.1 Source

Although tweets take up very little computer memory, at just 140 characters, the sheer quantity of tweets on twitter makes analyzing data on a large scale a monumental task. In 2013, Twitter published about 500 million tweets per day, and that number has increased since then. In this study we will take a small subset of twitter data of 1.6 million tweets from May 2009. These tweets were originally gathered by Stanford researchers interested in analyzing user sentiment through natural language processing [3]. In this study, we will use these tweets, their authors (usernames) and the hashtags they contain to create and analyze communities on twitter and users that belong to those communities.

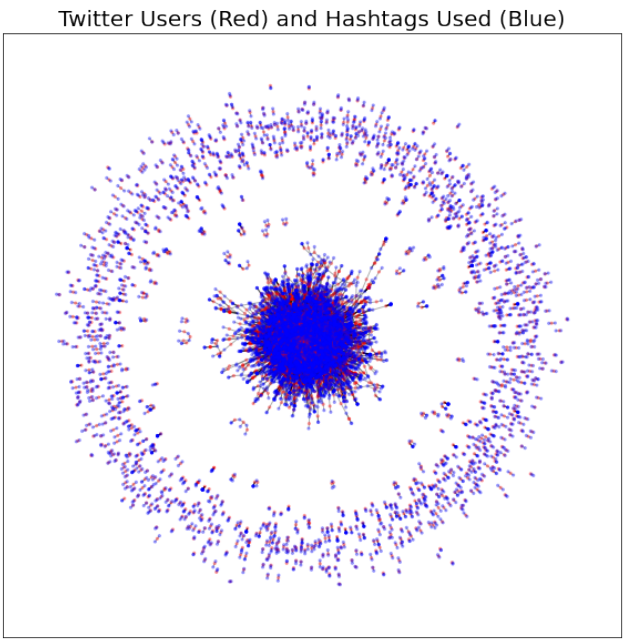

Figure 1: Complete Bipartite Graph of Twitter Users and Hashtags, with edges between users and hashtags that they use

2.2 Data Preparation

To prepare the data, we first iterate through each tweet in the dataset and create a set of all hashtags used and of all users. Since any string can be used as a hashtag, we need to discard hashtags that are only used by two or fewer users. This narrows the number of hashtags and users we need to consider somewhat, making computations easier and removing hashtags that are nonsensical or too specific to be interesting at a community level.

Now, with our pared-down sets of users and tweets, we can create a network with the data. First, we create a bipartite network with two sets of nodes, U for users and H for hashtags, where an edge exists between a user and a hashtag if that user used that hashtag in at least one tweet. At this point, we have a bipartite network with one large connected component and several smaller components, with a total of 15,157 nodes combined. See Figure 1. From there, we can isolate the giant connected component, discarding the smaller components, leaving us with 10,859 nodes. See Figure 2.

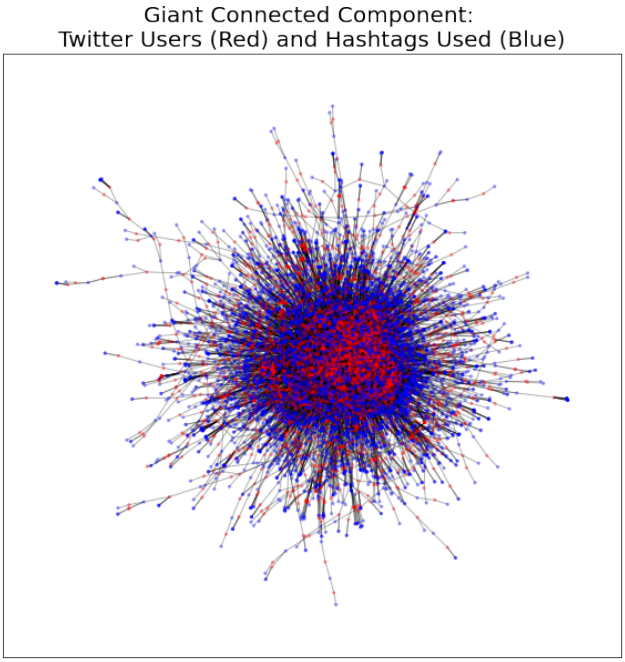

Figure 2: Complete Bipartite Graph of Twitter Users and Hashtags, with edges between users and hashtags that they use

3 Methodology

3.1 Finding Hashtag Communities

With our connected bipartite graph, we can proceed with finding communities within the network. First, we perform a unipartite projection of the bipartite network onto the hashtag nodes. This results in a connected network where every node is a hashtag, and two nodes are connected if a user used both hashtags in their tweets.

To focus on the most important hashtags, we can filter this network further by removing any hashtag in the network that was used less than 5 times total in the connected graph. This removes many hashtags that users used for hyper-specific reasons, hashtags with typos in them, and hashtags that are nonsensical. It also serves to make the network smaller and easier to process and analyze qualitatively. After discarding nodes left disconnected from this filtering process, we are left with 375 nodes in the network, each representing a popular hashtag. See Figure 3.

Now, we can perform community-detection on this network, with the aim of each hashtag community sharing relevance. We use the Louvain method for community detection, which optimizes the modularity Q of the sets of communities c in the graph G [2].

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{i,j} - \frac{1}{2m} \right] \delta(c_i, c_j) \tag{1}$$

where A is the adjacency matrix of G, m is the number of edges in the network, and $\delta(c_i, c_j)$ is the kronecker delta, which takes a value of 1 when the community of node i is

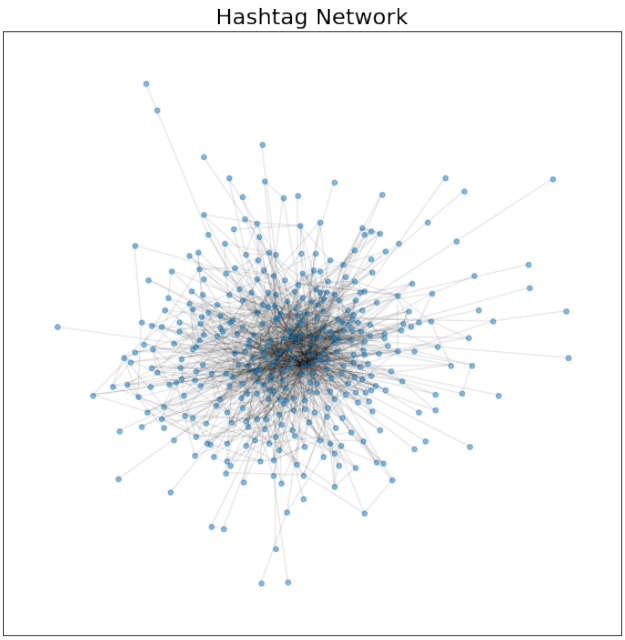

Figure 3: Unipartite projection of connected bipartite graph onto the hashtag nodes. Each node here represents a hashtag, and two nodes are connected if a user used both hashtags in their tweets.

the same as the community of node j and 0 otherwise. This method as implemented by Thomas Aynaud returns a partition of G with an optimal number of communities [1]. In this case, the method returns 15 communities within the network. See Figure 4. Now, because hashtags are complicated an language based, our best option is to evaluate the hashtag communities found using the Louvain method by hand. See Table 1.

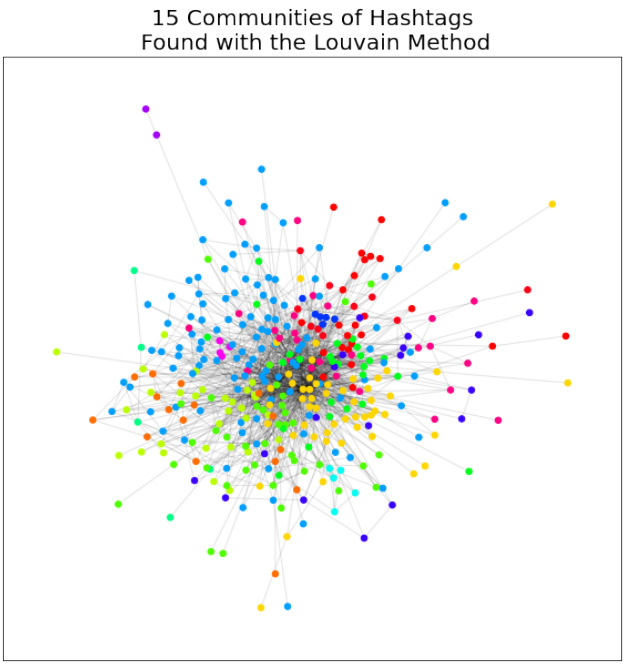

Figure 4: Communities found using the Louvain Method. Each community is represented by a different color. Note that some communities of hashtags are very small, while others are large.

Louvain-Detected Communities within Hashtag Network					
ID	Best Name	Size	Correct Hashtags	Incorrect Hashtags	Score
1	Iran Election	23	#tehran #iranelection #iran #iran9 #iran	#clothdiapers #eco #music4good #green	0.8
2	American Sports	16	#nba #Lakers #magic #redwings #hockey #nhl	#up #pens #android	0.85
3	Follow Friday and other Twitter Trends	53	#followfriday #FF #twittertakeover #samesexsunday #elevensestime #ecomonday #writechat #justsayin #followers	#otalia #dollhouse #spymaster #mallu #Firefox	0.75
4	American Music	30	#30secondstomars #marsiscoming #gokeyisadouche	#NEWO	0.4
5	Interpersonal Relationships	38	#3turnoffwords #myweakness #whocangetit #iconfess #liesgirlstell	#trance #140conf #eu09 #15 #lofnotc	0.6
6	Partying	26	#yaymen #twpp #vegas	#lost #teaparty #tech	0.2
7	None	5		#IPL #peterfacinelli #swineflu	0
8	Oh Yes They Did	5	#ontd #ohyeswedid #ohnotheydidnt	#fuqtwitter	0.9
9	Tech	112	#opensource #java #google #fb #twitter #gmail #Adobe #Apple #app #WWDC	#BBQ #art #sad #wine	0.9
10	Chuck (Show)	6	#Chuck #chuckeu #savechuck		1
11	Web Design	18	#squarespace #wordcamp #wcchicago #wordpress #jquery #blog #webdesign	#golf #drupal	0.8
12	A State of Trance (album)	2	#asot400 #ASOT400		1
13	Short Stack (band)	4	#shortstack #andyclemmensen #bradiewebb #shaundiveney		1
14	Television	26	#startrek #apprentice #pixar #maxout #nascar #rugby #starwars	#e3 #blogchat #facup	0.4
15	European Sports	11	#tennis #PakCricket #cricket #t20 #pakistan #frenchopen #cycling #federer	#windows	0.85

Table 1: Summary of Communities found in the Hashtag network using the Louvain Method. 14 of 15 of these communities had identifiable commonalities between hashtags. Several including 2, 3, 9, and 11 had nontrivial sizes and high qualitative relatedness as judged by the author.

3.2 Analyzing Community Membership of Users

Now that we have established communities of hashtags, we can analyze the users that use those hashtags (and therefore belong to those communities). Looking at the distribution of users that belong to n communities, we see that the great majority of users only belong to 1 of the 15 found communities, and the distribution decreases as the number of communities increases. See Figure 5.

To analyze which communities users are most likely to be a part of, we can plot the distribution of users in each community along with the number of hashtags that define each community. We see that the distributions are fairly similar, except in classes 3 and 8, which are the most popular communities. In Figure 6, we see that Community 3, "Follow Friday and other Twitter Trends" has a relatively small number of hashtags compared to the proportions of users that use them, reflecting the fact that this community appeals to a wide audience, while Community 6, "Partying", has a higher proportion of hashtags than members, perhaps indicating that the community has less established hashtags and less broad appeal.

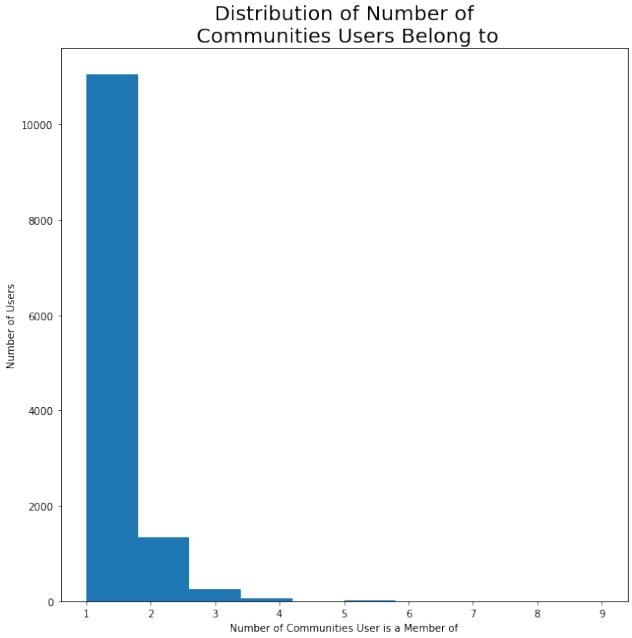

Figure 5: Histogram of Number of Communities users are members of.

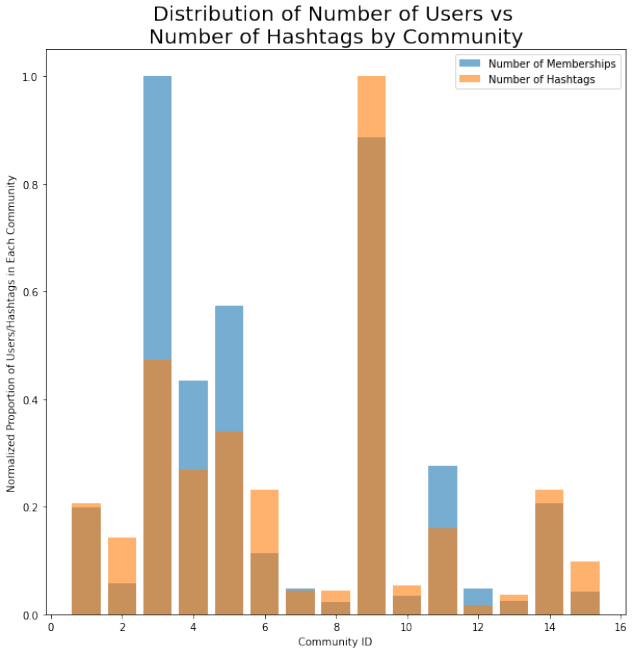

Figure 6: Distribution of proportion of users in each community and proportion of hashtags contained in each community.

4 Discussion and Conclusion

Our findings in this study show that communities of hashtags exist on Twitter and expose the nature of these communities. First, in view of Table 1, we see that hashtag communities are transient and news-cycle driven. Communities about the election in Iran, the TV show Chuck, the band Short Stack, and Tech were clearly driven by events in the news cycle during the week that these tweets were written in 2009. We also see that in semi-permanent communities, such as Tech or Follow Friday and other Twitter Trends that current events impact the hashtags included in each community during each week. For example, the week when these tweets were written was around the time the World-Wide-Developers conference in San Francisco, leading to a surge of hashtags relating to software, San Francisco, and Apple in the Tech community.

An interesting trend we see in the types of communities that emerge from the method presented here is that none of the communities seem to represent a dominant viewpoint on the topic they address. Because hashtags are generally categorical rather than statements of support or disapproval, hashtag communities are likely to be less homogeneous than, say, groups of users on Facebook. In this way, a conscientious Twitter user is more likely to gain a full, diverse understanding of an issue by searching for the hashtags associated with it, rather than scrolling through the tweets of the users that the user follows. Hashtags themselves can be a defense against polarization and misinformation on social media because diverse people with diverse viewpoints use the same hashtags to talk about the same issues.

If we look at the distributions of users in communities, we see that most users are members of one or just a few communities, at least at a given point in time. One of the interesting features and potential issues with this study is that the data was all recorded in a short time span. This gives us an almost instantaneous view of the hashtag communities; a view of what each community was like at each point in time. As we can see in Figure 5, there are relatively few users involved in more than one community at a time, which also supports the idea that the communities on Twitter and their users are transient, always changing in time.

For better or worse, Twitter is a very complicated and sometimes fickle social network. This accounts for at least some of the errors (uninterpretable classifications) that the Louvain Method made while creating hashtag communities. These errors include categorizing hashtags with no apparent meaning, creating community 7 with no apparent similarities, and classifying #windows with the European Sports instead of with Tech.

Despite these difficulties, however, it is clear that the Louvain method is capable of producing generally cohesive and useful communities of hashtags, which give insight into the types of communities that operate on Twitter.

4.1 Further Work

To better understand the relationships between hashtags used and the users that use them, a future study should consider creating a directed network of followers to better understand how the use of hashtags spread. This network could be compared to the hashtag network presented here to show the correlation between the communities users belong to and the communities that the people they follow belong to.

References

[1] T. Aynaud. Community detection for networkx's documentation.

[2] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Generalized louvain method for community detection in large networks. In *2011 11th international conference on intelligent systems design and applications*, pages 88–93. IEEE, 2011.

[3] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.