Beyond Keywords: Tracking the evolution of conversational clusters in social media

No author given

The potential of social media to give insight into the dynamic evolution of public conversations, and into their reactive and constitutive role in political activities, has to date been underdeveloped. While topic modeling can give static insight into the structure of a conversation, and keyword volume tracking can show how engagement with a specific idea varies over time, there is need for a method of analysis able to understand how conversations about societal values evolve and react to events in the world, incorporating new ideas and relating them to existing themes. In this paper, we propose a method for analyzing social media messages that highlights how world events become connected to existing conversations. This approach has applicability to the study of framing processes, and may reveal how contentious groups connect evolving events to their larger narratives.

1. MOTIVATION

Social media suggests a tantalizing prospect for researchers seeking to understand how newsworthy events influence public and political conversations. Messages on platforms such as Twitter and Facebook represent a high volume sample of the national conversation in near real time, and with the right handling[1] can give insight into events such as elections, as demonstrated

---

[1] The primary issues involve controlling for demographics. For examples of methods for handling demographic concerns with social media data, see Bail (2015) and McCormick,

by Huberty (2013) and Tumasjan, Sprenger, Sandner, and Welpe (2010); or processes such as social movements, as demonstrated by Agarwal, Bennett, Johnson, and Walker (2014) and DiGrazia (2015). Standard methods of social media analysis use keyword tracking and sentiment analysis to attempt to understand the range of perceptions surrounding these events. While helpful for basic situational awareness, these methods do not help us understand how a set of narratives compete to interpret and frame an issue for action.
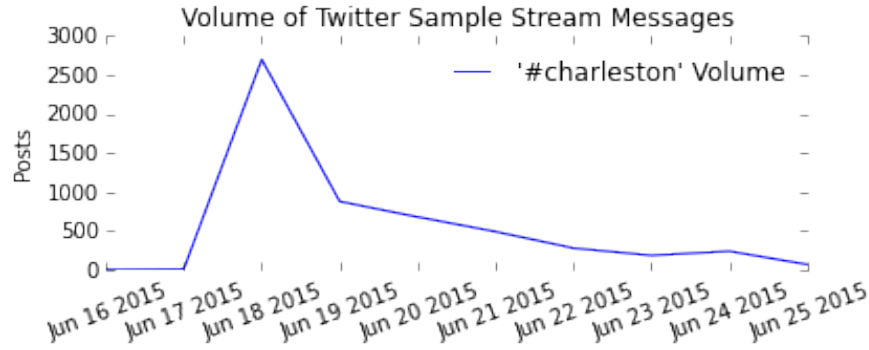

Figure 1. Standard methods of social media analysis include keyword volume tracking (shown here), sentiment analysis, and supervised categorization.

For example, if we are interested in understanding the national conversation in reaction to the shooting at Emanuel AME church in Charleston, South Carolina in June of 2015, we could plot a timeseries of the volume of tweets containing the hashtag `#charleston`, as seen in Figure 1. This tells us something about how many people are engaging with the topic, but little about the ways they are interpreting the event or how they connect the event to their preexisting ideas about gun violence.

Lee, Cesare, Shojaie, and Spiro (2015).

2

Alternate methods include assessing the relative 'positive' or 'negative' sentiment present in these tweets, or using a supervised learning categorizer to group messages according to preconceived ideas about their contents, as demonstrated by Becker, Naaman, and Gravano (2011), Ritter, Cherry, and Dolan (2010), and Zubiaga, Spina, Fresno, and Martínez (2011). Such methods can give aggregated insight into the sentiment expressed, but not into the ways the events are being framed within existing conversations.

Because they look at individual messages and not at the relationships between ideas within messages, these techniques are unable to infer from the data coherent patterns of thought that signify interpretation of the events' deeper meanings. Interpretation depends upon making connections between the events as they happen and other concepts in the public discourse. One way to measure these connections is to look at a network of co-citations surrounding our topic of interest, as has been demonstrated by Cogan and Andrews (2012) and Smith and Rainie (2014). This technique represents hashtags as nodes on a network, and each message containing two or more hashtags contributes to the weight of an edge between these nodes. This type of analysis is helpful in that it helps us begin to understand the structure of the discourse. If we perform k-clique clustering on the network, we can see which sets of connections form coherent and conversations, as seen in Figure 2.

In this example there seem to be two distinct conversations happening with regards to the event: the first a description of the shooting itself and
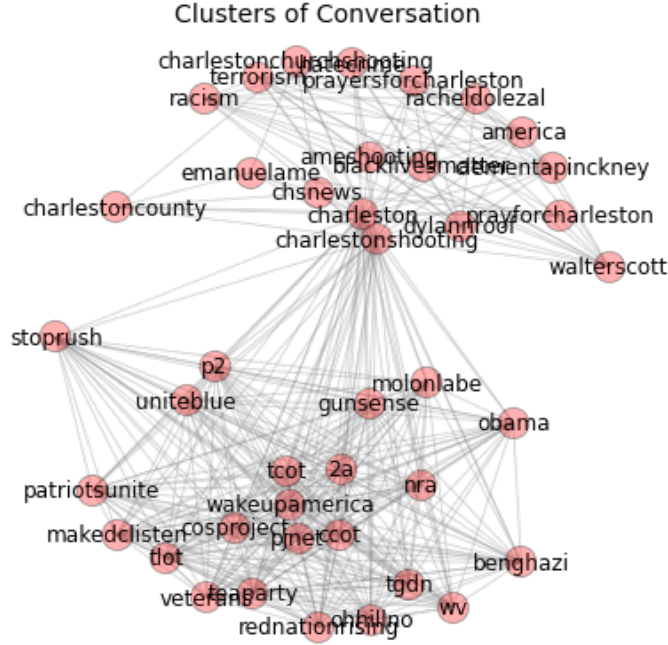
Figure 2. K-clique clustering reveals coherent structures in the hashtag co-citation network that can serve as proxies for conversations in the larger national discourse.

the human elements of the tragedy. A second conversation focuses on the larger national-scale political conflicts to which the event points. While each of the conversations is motivated by the same event, they are distinct from one another in the language they use and in the connections they draw.

1.1 The contributions of this paper

We may hypothesize that how these conversations develop over time will influence the social and political response to the event. In this paper we will explore methods of identifying and tracking the development of these conversations over time. We show how these conversation clusters can be identified

and visualized, exploring how certain elements form the core of a conversation, while other elements circulate at the periphery. We then demonstrate a method for tracking the evolution of these conversational clusters over time, by relating clusters identified on one day with clusters emerging on subsequent days. This allows us to qualitatively see what new concepts are being included into the discussion, and quantitatively track engagement in these conversations, as distinct from mere references to keywords.

Due to the computation-intensive nature of this analysis, we chose to implement the data manipulation algorithms in both Python and Unicage shell scripts,[2] for prototyping and speed of execution, respectively. Descriptions of these scripts can be found in the appendices, along with performance comparisons between the two languages.

2. IDENTIFYING CONVERSATION CLUSTERS

The image in Figure 2 is based upon network closeness between hashtags. Each of the hashtags present in the dataset forms a node in this network, and the relative strength of edges depends upon the number of times the pair occur together in a tweet, their 'co-occurrence', using the method of Marres and Gerlitz (2014).

The clusters themselves are then defined by k-clique community detection algorithms implemented in the COS Parallel library, developed by Gregori, Lenzini, and Mainardi (2013) and their use is demonstrated in the appen-

---

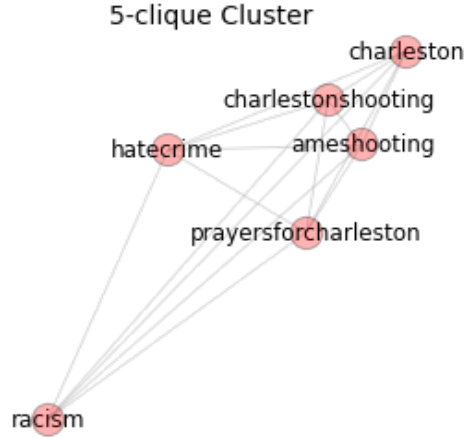[2]For a description of Unicage development tools, see Tounaka (2013)

dices.

*Figure 3. Clusters with higher **k** value are smaller and more tightly connected, representing a more coherent or focused conversation.*

Every node in a cluster must be able to participate in at least one fully connected subgroup of size **k** with the other members of the group. Thus, the metric **k** determines how strict we are about closeness between keywords when defining the boundaries of a particular cluster. For example, high values of **k** would impose strict requirements for interconnectedness between elements of an identified conversation, leading to a smaller, more coherent identified conversation, as seen in Figure 3.

On the other hand, smaller values of **k** are less stringent about the requirements of connectivity they put on the elements in the cluster, leading to a larger, more loosely coupled group, as seen in Figure 4.
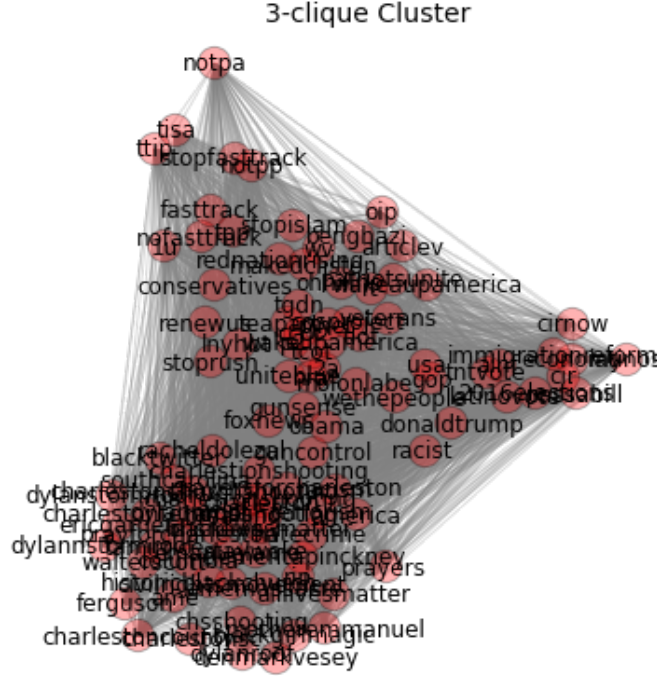
*Figure 4. Clusters with lower **k** value are larger and less less tightly connected, representing more diffuse conversation. They may have smaller clusters of conversation within them.*

3.   REPRESENTING CONVERSATIONAL CLUSTERS AS NESTED SETS

Tight conversational clusters (high `k`) must necessarily be contained within larger clusters with less stringent connection requirements (low `k`). Performing clustering along a range of `k` values allows us to place a specific conversation in context of the larger discourse. It becomes helpful to represent these clusters as nested sets, as seen in Figure 5, ignoring the node and edge construction of the network diagram in favor of something which allows us

to observe the nested relationships the conversations have with one another.

In this representation, we are able to observe the tightly clustered 5-clique conversation in context of the 4-clique conversation it inhabits, and the neighboring 4-clique conversations that together inhabit the larger discourse.

4. TRACKING CONVERSATIONS CHRONOLOGICALLY

In order to track how elements of conversation weave into and out of the general discourse, we need to be able to interpret how conversational clusters identified at one point in time relate to those in subsequent intervals. We can do this in one of two ways.

The first method is to track the volume of co-citations identified in the various conversational clusters identified on the first day of the analysis, as it changes over subsequent days. This indicates how well the connections made on the first day maintain their relevance in the larger conversation. Figure 6 shows how the connections made in the conversational clusters shown in Figure 5 fall in volume over the 10 days subsequent to the initial event, paralleling the decay in pure keyword volume seen in Figure 1.

The second method for tracking conversation volume over time takes into account the changes that happen within the conversation itself. The fundamental assumption in this analysis is that while the words and connections present in a conversation change, they do so incrementally in such a way as to allow for matching conversations during one time period with those in the

8

prayforcharleston
charlestoncounty
charleston
charlestonshooting
emanuelame
chsnews
charlestonchurchshooting

racheldolezal
america
terrorism
blacklivesmatter
clementapinckney
walterscott
dylannroof

hatecrime
charleston
prayersforcharleston
charlestonshooting
racism
ameshooting

charleston
2a
wakeupamerica
molonlabe
ccot
p2
obama
pjnet
stoprush
wv
tlot
nra
tcot
makedclisten
benghazi
veterans
teaparty
patriotsunite
charlestonshooting
ohhillno
cosproject
tgdn
gunsense
rednationrising
uniteblue

Figure 5. Converting networks to nested sets based upon k-clique clustering simplifies presentation and analysis of various levels of conversation.
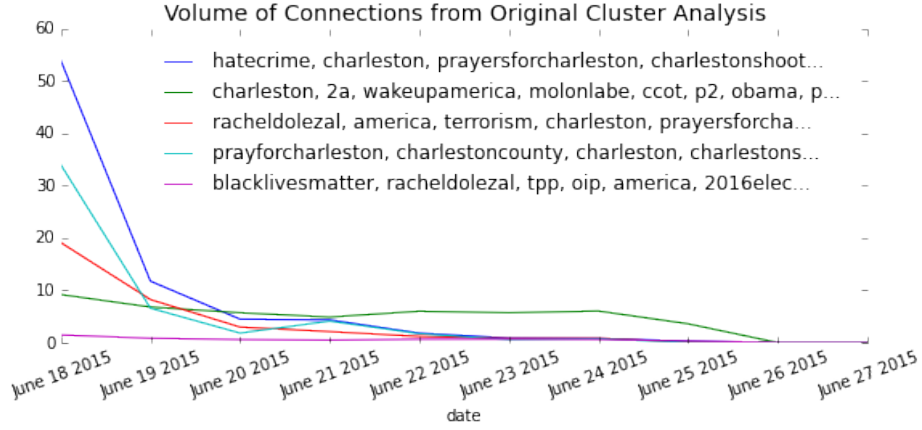
Figure 6. Tracking the volume of connections made in a single day's clusters (e.g. co-citations) reveals how the specific analogies made immediately after the event maintain their relevance.

immediately following time period.

Palla, Barabási, and Vicsek (2007) discuss how communities of individuals develop over time and change. We can use the same techniques to track continuity of conversational clusters. The most basic way to do this is to count the fraction of elements of a conversational cluster at time 1 that are present in each conversational cluster at time 2, and use this fraction as the likelihood that each cluster at time 2 is an extension or contraction of the time 1 cluster in question. From this we can construct a transition matrix relating conversational clusters at time 1 with clusters at time 2, as seen in Table 1.

To improve our estimates, we can take advantage of the fact that clusters that correspond from time 1 to time 2 will participate in a larger cluster that emerges if we perform our clustering algorithm on the union of all edges

10

Table 1. A transition matrix contains the likelihood that a cluster at one timeperiod (rows) corresponds to a cluster in the subsequent timeperiod (columns).

Cluster ID	t2-cl1	t2-cl2	t2-cl3	. . .
t1-cl1	0.68	0.2	0.0	. . .
t1-cl2	0.0	0.85	0.03	. . .
t1-cl3	0.13	0.04	0.45	. . .
. . .	. . .	. . .	. . .	. . .

from the networks at time 1 and time 2. This reduces the number of possible pairings between days, yielding more specificity in our intra-day transition matrices.

These transition matrices can be used to infer how clusters present in the first day's analysis correspond to clusters in the second day, and so forth. These are visualized as a set of nested cluster diagrams, with traces linking likely clusters together, as seen in Figure 7. Heavier traces between clusters imply more confidence in the transition between the two sets.

The volume of messages forming the edges of each cluster are shown plotted day by day in Figure 8. As the linkage between subsequent clusters is probabilistic, in this plot so to are the links connecting volume measures, with the weight of each link proportional to its likelihood. In the present example, we can see how following expressions of the event itself, the conversation evolves into a discussion of the relationship between the violence and the ongoing issue of racism and its connection with the confederate flag. Parallel to and as an undercurrent of these conversations is the ongoing discussion of
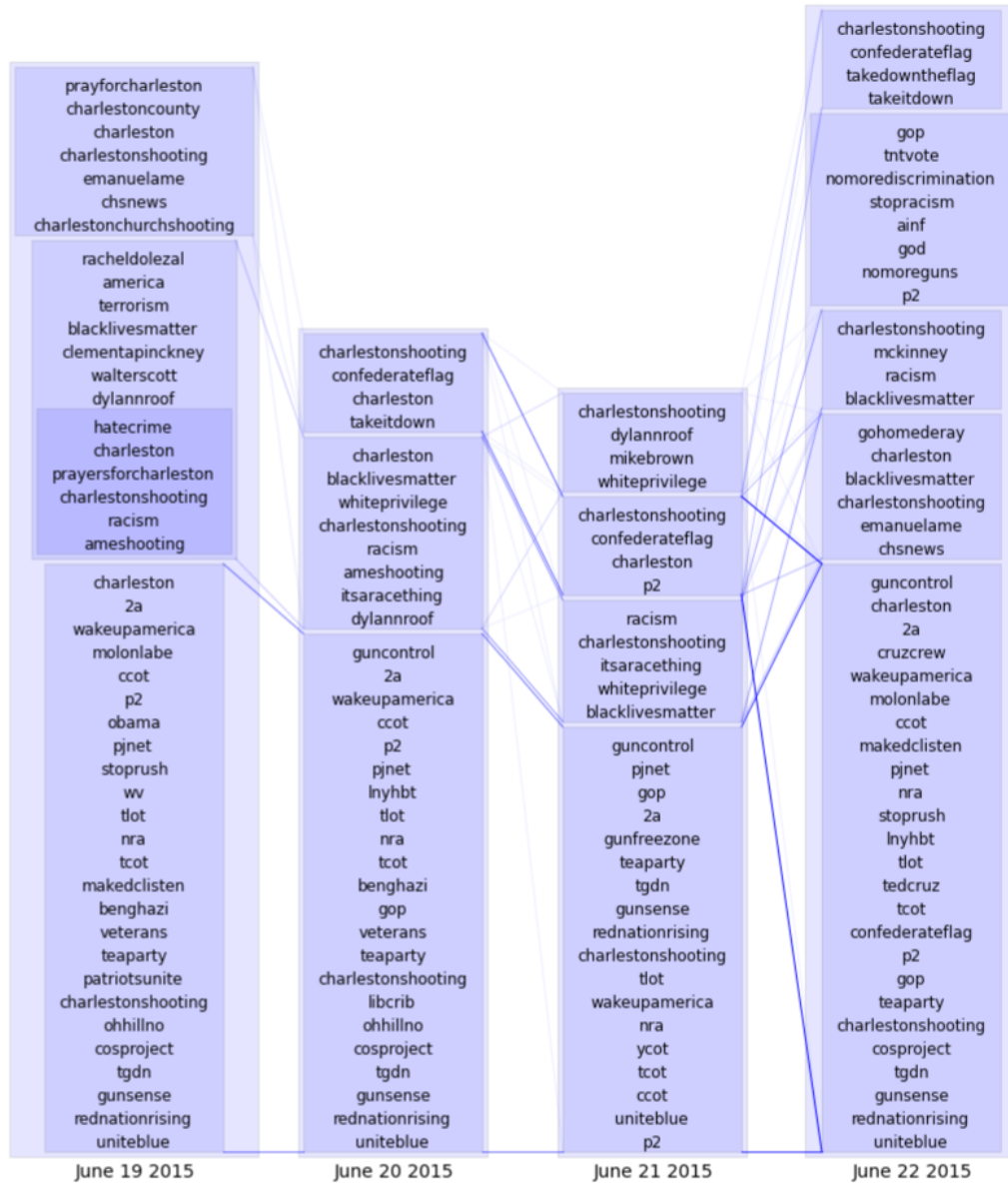
Figure 7. Weighted traces connect conversation clusters for four days following the shooting. Conversations change over time, as certain components fall out of the cluster, and other components are added.

Figure 8. Plotting over time the mean connection volume in in each conversational cluster, showing links according to the probabilistic transition matrix. Clusters show relative consistency, with new terms occasionally raising participation.

gun control.

5. CONCLUSION

The utility of social media analysis for sociological research can be extended well beyond the practice of tracking keyword volume, net post sentiment, or supervised classification. In particular, tracking conversational clusters within a network of hashtag co-citations can give both structural understanding of a conversation, and insight into how it develops over time.

These tools can prove helpful to sociologists interested in using social

media to understand how world events are framed within the context of existing conversations.

Follow-on work to this paper could attempt to separate the various conversational clusters according to the groups engaged in them, possibly using information about the Twitter connection graph to understand if certain conversations propagate through topologically separate subgraphs, or if multiple conversations occur simultaneously within the same interconnected communities. Such research would have obvious impact on our understanding of framing, polarization, and the formation of group values.

6. AUTHOR'S NOTE

The appendices to this paper contain all of the code needed to collect necessary data, generate the analysis, and and produce visualizations found herein:

Appendix A Cluster Identification and Transition Analysis in Python

Appendix B Cluster Identification and Transition Analysis in Unicage

Appendix C Performance Comparison Between Python and Unicage Examples

Appendix D Data Collection Scripts

Appendix E Visualizations

The full set of scripts, and associated documentation, can be found at `www.github.com\ removed to anonymize`.

REFERENCES

Agarwal, S. D., Bennett, W. L., Johnson, C. N., and Walker, S. (2014), "A model of crowd enabled organization: Theory and methods for understanding the role of twitter in the occupy protests," *International Journal of Communication*, 8, 646–672.

Bail, C. A. (2015), "Taming Big Data: Using App Technology to Study Organizational Behavior on Social Media," *Sociological Methods & Research*, 0049124115587825–.

Becker, H., Naaman, M., and Gravano, L. (2011), "Beyond Trending Topics: Real-World Event Identification on Twitter.," *ICWSM*.

Cogan, P. and Andrews, M. (2012), "Reconstruction and analysis of twitter conversation graphs," *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*.

DiGrazia, J. (2015), "Using Internet Search Data to Produce State-level Measures: The Case of Tea Party Mobilization," *Sociological Methods & Research*, 0049124115610348–.

Gregori, E., Lenzini, L., and Mainardi, S. (2013), "Parallel k-clique commu-

nity detection on large-scale networks," *IEEE Transactions on Parallel and Distributed Systems*, 24(8).

Huberty, M. (2013), "Multi-cycle forecasting of congressional elections with social media," *Proceedings of the 2nd workshop on Politics, Elections and Data.*

Marres, N. and Gerlitz, C. (2014), "Interface Methods: Renegotiating relations between digital research, STS and Sociology," unpublished manuscript.

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E. S. (2015), "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing," *Sociological Methods & Research*, 0049124115605339–.

Palla, G., Barabási, A., and Vicsek, T. (2007), "Quantifying social group evolution," *Nature.*

Ritter, A., Cherry, C., and Dolan, B. (2010), "Unsupervised modeling of twitter conversations," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 172–180.

Smith, M. and Rainie, L. (2014), "Mapping twitter topic networks: From polarized crowds to community clusters," unpublished manuscript.

Tounaka, N. (2013), "How to Analyze 50 Billion Records in Less than a Second without Hadoop or Big Iron,".

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010), "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.," *ICWSM.*

Zubiaga, A., Spina, D., Fresno, V., and Martínez, R. (2011), "Classifying trending topics: a typology of conversation triggers on twitter," *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2461–2464.