Beyond Keywords: Tracking the development of conversations on social media through linked, nested hashtag clusters

James Houghton, Michael Siegel: Massachusetts Institute of Technology
Nobuaki Tounaka, Buyanjargal Shirnen, Daisuke Nakagawa, Kazutaka Nakamura: USP-Labs

Motivation:

In seeking to know how newsworthy events influence public and political conversations, researchers benefit from data available on Social Media. Messages on platforms such as Twitter and Facebook represent a high volume sample[1] of the national conversation in near real time. 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.

For example, if we are interested in understanding the national conversation in reaction to the shooting at Emanuel AME church in Charleston, South Carolina, we could plot a timeseries of the volume of tweets containing the hashtag #charleston, as seen in Figure 1.

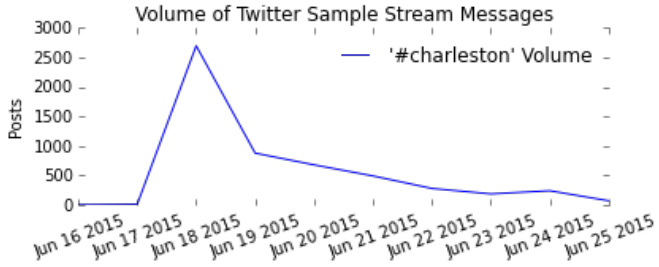

Figure 1: Standard methods of social media analysis include keyword volume tracking (shown here), sentiment analysis, and supervised categorization

[1] Caution must be taken as SM samples are not demographically representative.

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.(Becker, Naaman, & Gravano, 2011; Ritter, Cherry, & Dolan, 2010; Tumasjan, Sprenger, Sandner, & Welpe, 2010; Zubiaga, Spina, Fresno, & Martínez, 2011)

These techniques, however, 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 as a network of co-citations surrounding our topic of interest. (Cogan & Andrews, 2012; Smith & Rainie, 2014) Representing hashtags as nodes on a network, each shared message 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 are two distinct conversations happening with regards to the event – the first a description of the shooting itself and the human elements of

the tragedy. A second conversation focuses on the larger national-scale political conflicts to which the event points. While

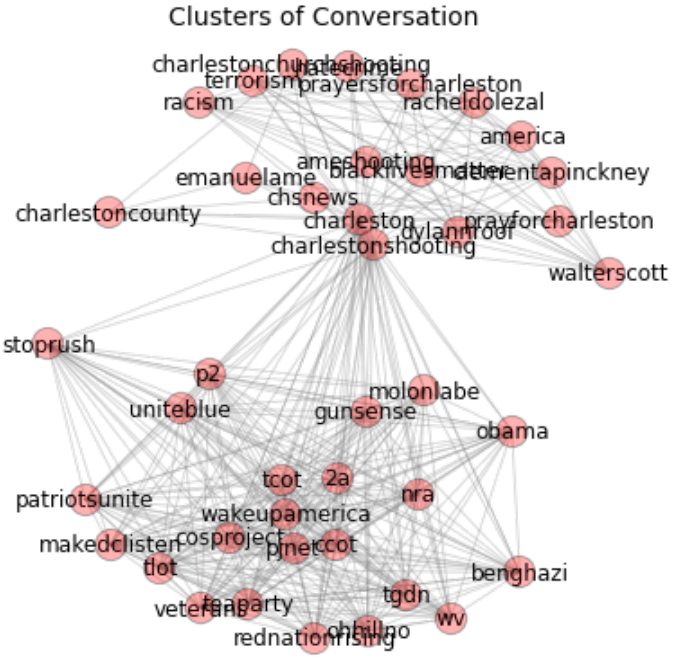

Figure 2: K-clique clustering reveals coherent structures in the hashtag co-citation network that can serve as proxies for conversations in the larger national discourse.

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 make.

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.

Identifying Conversation Groups:

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'. (Marres & Gerlitz, n.d.)

Scripts for performing this data handling and network construction using both Python and Unicage(Tounaka, 2013) shell scripts can be found in the appendicies, along with performance comparisons between the two languages.

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

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.

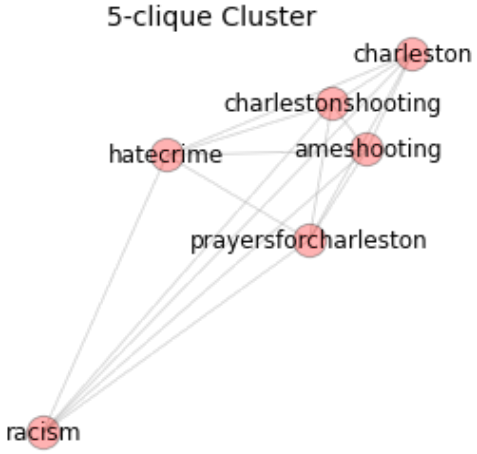

Figure 3: Clusters with higher 'k' value are smaller and more tightly connected, representing a more coherent or focused conversation.

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.

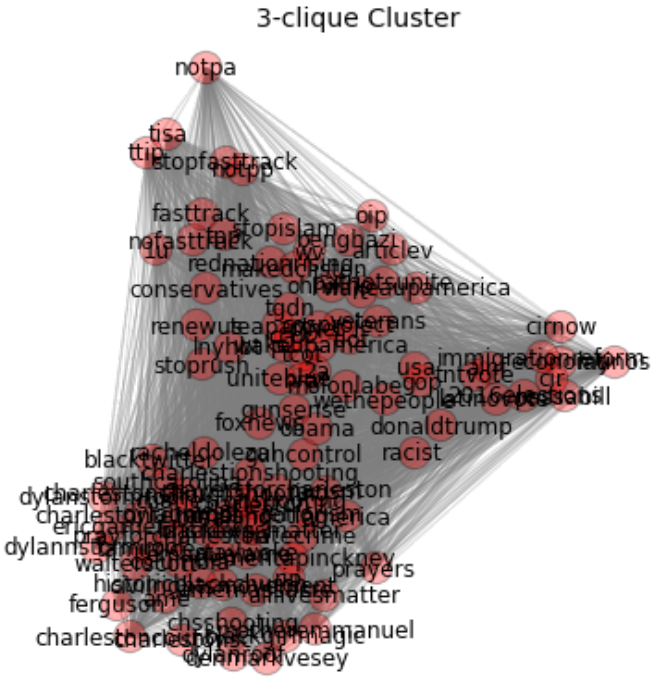

Figure 4: Clusters with lower 'k' value are larger and less less tightly connected, representing more diffuse conversation. They may have smaller clusters of conversation within them.

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.

Figure 5: Converting networks to nested sets based upon k-clique clustering simplifies presentation and analysis of various levels of conversation.

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 above 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 following time period.

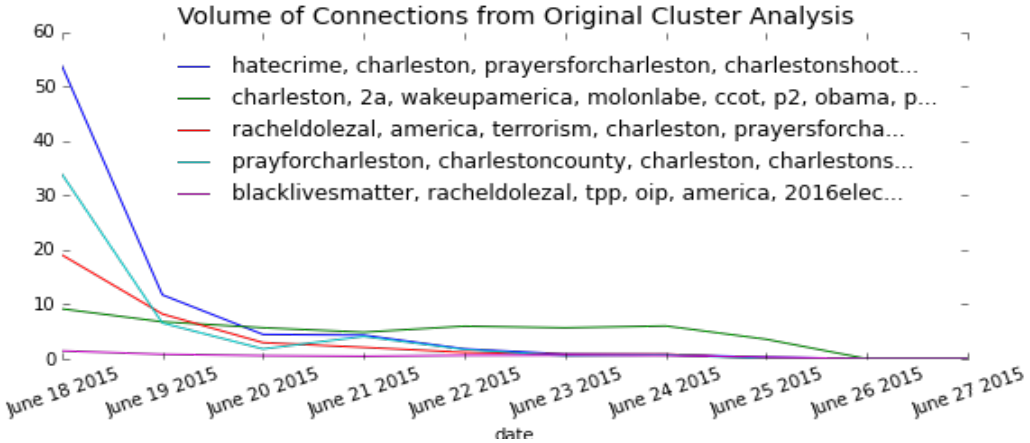

Figure 6: Tracking the volume of connections made in a single day's clusters (ie, co-citations) reveals how the specific analogies made immediately after the event maintain their relevance.

Palla et al(Palla, Barabási, & Vicsek, 2007) have discussed 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 0 that are present in each conversational cluster at time 1, and use this fraction as the likelihood that each cluster at time 1 is an extension or contraction of the time 0 cluster in question. From this we can construct a transition matrix relating conversational clusters at time 0 with clusters at time 1, as seen in Table 1.

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	20150619_k4_t5_i30	20150619_k4_t5_i159	20150619_k4_t5_i103	20150619_k3_t5_i13
20150618_k4_t5_i15	0.62069	0	0	0
20150618_k4_t5_i54	0	0.4	0	0
20150618_k4_t5_i140	0	0	0.222222	0
20150618_k3_t5_i4	0	0	0	0.366667

To improve our estimates, we can take advantage of the fact that clusters that correspond from time 0 to time 1 will participate in a larger cluster that emerges if we perform our clustering algorithm on the union of all edges from the networks at time 0 and time 1. This reduces the number of possible pairings between days, yielding

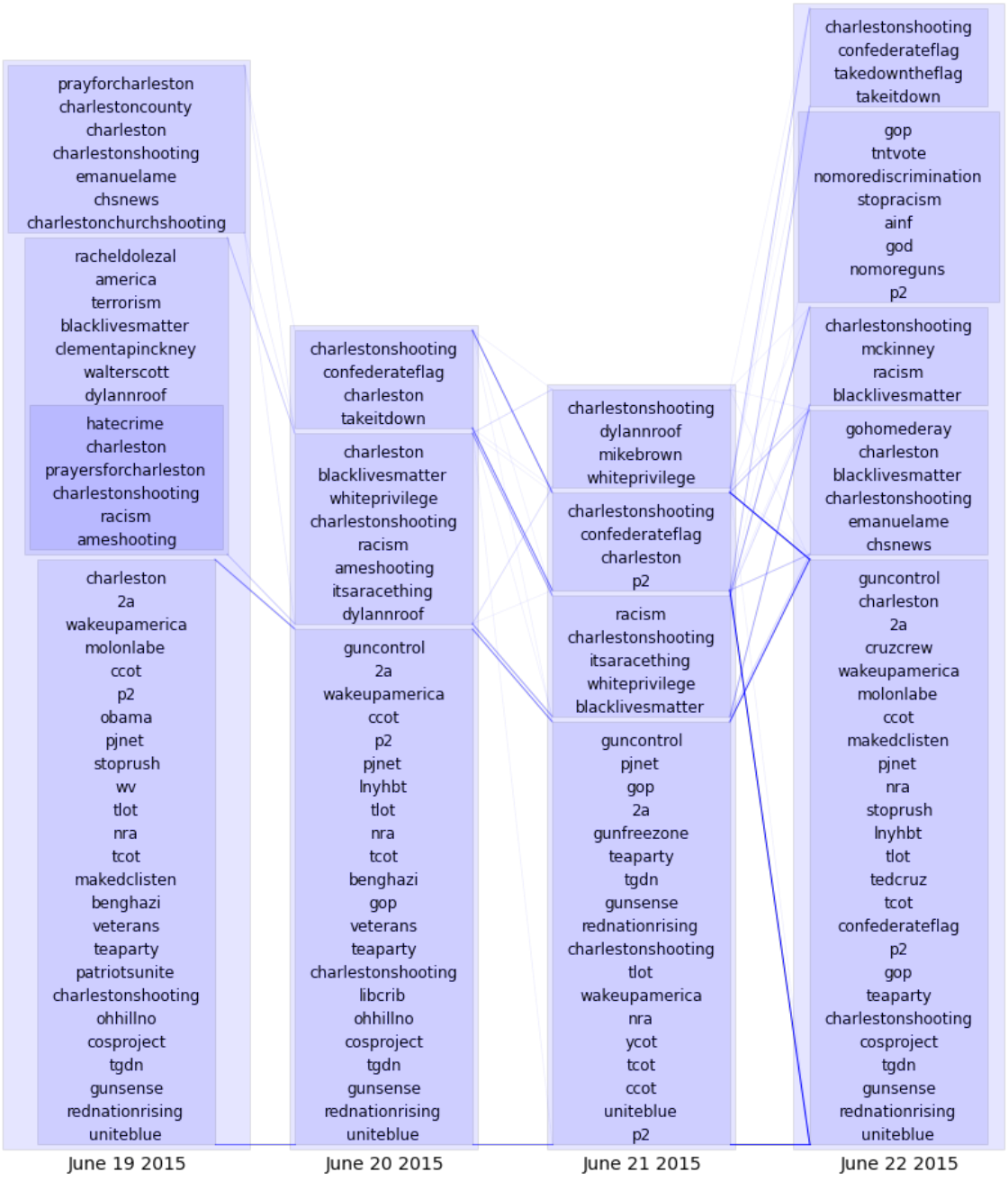

Figure 7: Weighted traces conect conversation clusters for four days following the shooting. Conversations change over time, as certain components fall out of the cluster, and other components are added.

more specificity in our intra-day transition matricies.

These transition matricies 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.

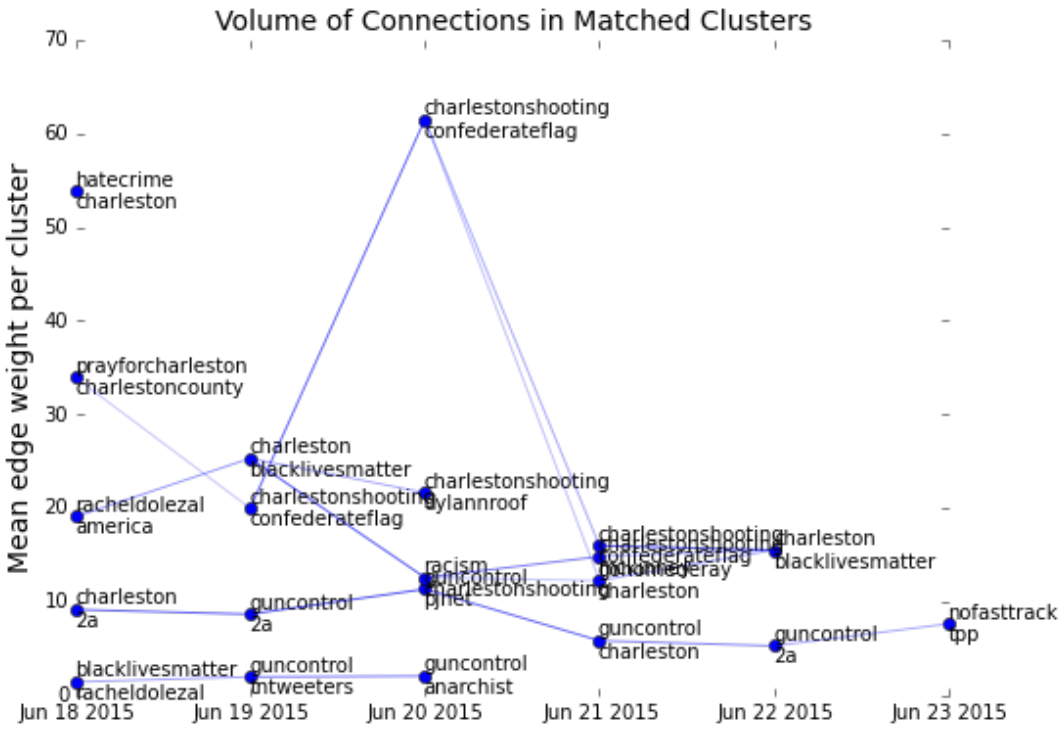

Figure 8: Plotting over time the mean connection volume in in each conversational cluster, showing links according to the probabilistic transition matrix. Clusters show relative consistency, with new terms occasionally raising participation.

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.

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.

Bibliography

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. *ICWSM*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2745Hila/3207

Cogan, P., & Andrews, M. (2012). Reconstruction and analysis of twitter conversation graphs. *Proceedings of the First ….* Retrieved from http://dl.acm.org/citation.cfm?id=2392626

Gregori, E., Lenzini, L., & Mainardi, S. (2013). Parallel k-clique community detection on large-scale networks. *Parallel and Distributed ….* Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6249683

Marres, N., & Gerlitz, C. (n.d.). *Interface Methods: Renegotiating relations between digital research, STS and Sociology* (No. 3). *CSISP Working Papers*. Retrieved from

http://uranus.gold.ac.uk/media/Marres__Gerlitz_workingpaper_3.pdf

Palla, G., Barabási, A., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*. Retrieved from http://www.nature.com/nature/journal/v446/n7136/abs/nature05670.html

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised modeling of twitter conversations. Retrieved from http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=16885300

Smith, M., & Rainie, L. (2014). Mapping twitter topic networks: From polarized crowds to community clusters. *Pew Research Internet ….* Retrieved from http://www.pewinternet.org/2014/02/20/part-2-conversational-archetypes-six-conversation-and-group-network-structures-in-twitter/

Tounaka, N. (2013). How to Analyze 50 Billion Records in Less than a Second without Hadoop or Big Iron. Retrieved from https://en.usp-lab.com/index.html

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852Predicting

Zubiaga, A., Spina, D., Fresno, V., & Martínez, R. (2011). Classifying trending topics: a typology of conversation triggers on twitter. *Proceedings of the 20th ACM ….* Retrieved from http://dl.acm.org/citation.cfm?id=2063992

Appendicies:

These appendices list the data handling and analysis code necessary to perform the conversation identification and tracking demonstrated in this paper. These scripts are also available on github at https://github.com/JamesPHoughton/twitter-cluster.

Appendix A: Cluster Identification and Transition Analysis in Python
Appendix B: Cluster Identification and Transition Analysis with Unicage Shell Scripting
Appendix C: Performance Comparison
Appendix D: Data Collection
Appendix E: Visualization