Memes Online: Extracted, Subtracted, Injected, and Recollected

Matthew P. Simmons, Lada A. Adamic, Eytan Adar
School of Information
University of Michigan
Ann Arbor, MI 48109, USA
{mpsimmon,ladamic,eadar}@umich.edu

Abstract

Social media is playing an increasingly vital role in information dissemination. But with dissemination being more distributed, content often makes multiple hops, and consequently has opportunity to change. In this paper we focus on content that should be changing the least, namely quoted text. We find changes to be frequent, with their likelihood depending on the authority of the copied source and the type of site that is copying. We uncover patterns in the rate of appearance of new variants, their length, and popularity, and develop a simple model that is able to capture them. These patterns are distinct from ones produced when all copies are made from the same source, suggesting that information is evolving as it is being processed collectively in online social media.

1 Introduction

Sherlock Holmes never uttered : “Elementary, my dear Watson”. He did say “elementary” and “my dear Watson” quite often, and at some point, someone concatenated the two. The increased fitness of this meme over the two original quotes helped it to propagate, eventually making it the most well known quote attributed (inaccurately) to Sherlock Holmes. Before the advent of online social media, such memes were difficult to trace. However, recent work has begun to do just that (Leskovec, Backstrom, and Kleinberg 2009).

Although we are gaining an understanding of how the popularity of memes waxes and wanes, it is not well understood how they change. It is likely that social media is not just facilitating the propagation of memes, but also inhibiting or enhancing their mutation rate. On the one hand, online social media, by being digital, make it easy for individuals to make perfect copies of information. On the other hand, by providing many diffusion channels and inundating individuals with more information then they could possibly consume, it creates an environment where mutations can confer swift advantage in spread. Some changes to information in this “telephone game” are benign, e.g, correcting a misspelling or abbreviating an overly lengthy quote. But others can change the content’s meaning in both intentional and unintentional ways. In a recent example (Copus 2011), one Twitter user tweeted,

> *“Street style shooting in Oxford Circus for ASOS and Diet Coke. Let me know if you’re around!!”*

Three minutes later, another tweeted *“Shooting in progress in Oxford Circus? What?”*. This then mutated to a form that was readily retweeted and even emailed within minutes: *“Shooting in progress in Oxford Circus, stay safe people.”*

When mutations like this occur they raise questions about the authority and reliability of information diffusion through social media. In this case, the correct information, that there was no gunman, was propagated within a half hour, but in other cases, changes in information persist. Gaining a better understanding of the conditions that influence the fidelity of information as it diffuses will aid in developing effective systems to assist users in determining the authority of the content they consume.

In this paper we take the first step to understanding how memes change online. We do this by looking for mutation in places where it should be occurring the least, namely quoted text. Our measurements comprise a baseline, with the expected rate of change for unquoted text being much higher. Quotes nominally ought to represent text copied from another source. However, this text can be *reframed*, by subtracting from or adding to the beginning or end of the quote, or *altered* with omissions and substitutions.

Using a large corpus of quoted and clustered text from the MemeTracker project (Leskovec, Backstrom, and Kleinberg 2009), we quantify the prevalence of changes in memes. We show that the type and likelihood of change depends on whether the quotation is written on a blog or a mainstream media site, and the popularity of the source that is being copied from. We further examine properties of quote variants, when they occur and how popular they are, and develop a simple copying model that can replicate the main features of the data. Finally, we provide further evidence that the observed patterns are a product of derivative copying, by showing that direct quotes of New York Times (NYT) articles made by bloggers fail to display the same characteristics.

2 Related Work

Information diffusion in the blogosphere has been the subject of much prior work. Typically the goal has been to infer the path of the information through the blogosphere, from observations of unique strings, such as URLs, being cited by multiple sources (Adar and Adamic 2005;

Gruhl et al. 2004). Identifying information flow has relied on timing (Leskovec, Adamic, and Huberman 2007; Kumar et al. 2005), content analysis (Fisher et al. 2008; Berendt and Subasic 2009; Gomez Rodriguez, Leskovec, and Krause 2010) as well as hybrid approaches combining link and content analysis (Nallapati et al. 2004). However, the focus has been on tracing the *same* piece of information, and not how this information changes. In contrast, the change in information is the primary focus of the present paper.

The MemeTracker (MT) dataset (Leskovec, Backstrom, and Kleinberg 2009), used in this paper, has enabled a series of analyses of patterns in online information diffusion. Gomez Rodriguez, Leskovec, and Krause (2010) use the MemeTracker data to develop and test an algorithm for inferring networks of information diffusion using only adoption timestamps for a number of assets diffusing over the network. Ennals, Trushkowsky, and Agosta (2010) use MemeTracker data to identify disputed claims on the web. Finally, Yang and Leskovec (2011) define a sequence of 6 distinctive patterns of information diffusion in online media using the same dataset. None so far have addressed the question of meme mutation.

Other research has focused not on how individual items diffuse in the blogosphere, but how information on a topic evolves over time. There has been particular interest in clustering temporally and thematically similar events in large news corpora (Shinyama, Sekine, and Sudo 2002; Balahur et al. 2009; Barzilay and Lee 2003), exploring the evolution of scientific topics over time (Hall, Jurafsky, and Manning 2008), and on tracing the evolution of themes (Mei and Zhai 2005). These studies have dealt with accumulation of information about an event or topic over time, which necessarily involves aggregation from multiple sources. They have not addressed the fidelity of information as it propagates away from a single source.

3 Datasets

For our analysis we focus on two datasets, each consisting of quoted text extracted from online news articles and weblog posts. The first is the MT dataset, and the other is a series of online news articles from the New York Times website (nytimes.com) and a large number of blog posts that refer to those articles.

3.1 MemeTracker

The MemeTracker dataset is a publicly available collection of quotes and links that were extracted during a nine month crawl of the blogosphere (from August 2008 to April 2009). In its raw form, the data consists of the quotes and links extracted from posts, as well as the address of the crawled post and the time it was crawled. In total it contains 210,999,824 quotations and 418,237,269 links from 96,608,034 different blog posts.

The MemeTracker data also contains a clustering, described in (Leskovec, Backstrom, and Kleinberg 2009), of the quoted strings (i.e., phrases or quotes) into "phrase clusters" based on string similarity. There are 71,568 phrase clusters containing 310,547 unique quotations extracted from 7,665,108 different blog posts and news articles. The clustered data contains the URL and time of each post that contains a quotation along with the phrase cluster that the quote belongs to and a classification of the post as belonging to either *mainstream media* or a *blogger*. To facilitate the tracking of the memes, we further augment the clusters with the hyperlinks present in the raw data. After integrating the hyperlink information, our working dataset shrinks slightly to 309,730 distinct quotes in 71,344 clusters extracted from 6.85 million blog posts over 415,000 distinct hosts. This leaves us with a sufficient amount of data to detect even subtle changes in the patterns of content that represent meme mutation.

3.2 BlogRunner

To augment our analysis we also captured data from the New York Times BlogRunner service. The service provides links to blog posts referencing NYT articles. For 90 days we recorded the top 20 "most blogged about" stories from the NYT and then crawled the BlogRunner links to capture the text of the blog entries discussing the stories. We gathered 8,091 blog posts relating to 125 distinct NYT news stories. We then processed the source text to extract the links and the quoted strings from the news stories and blog posts.

The data differs from MemeTracker in two important ways. First, the clustering applied to the NYT stories was not string–based as the MT clustering was. Instead each cluster consists of quotes in posts relating to a *single* NYT article. Second, because all blog entries captured directly refer to the source news story, this dataset represents only a single generation for potential mutation to occur. In contrast, the time scale and scope of the MemeTracker data makes it likely that some of the memes have gone through a number of intermediate steps in their propagation from the original source, with each generation susceptible to mutation.

3.3 Terminology and Filtering

Terminology We co-opt the following terminology from (Leskovec, Backstrom, and Kleinberg 2009):

phrase or **quote**: a distinct quoted string extracted from a source.

mention: occurrence of a phrase or quote in the data, typically in a blog post.

phrase cluster or simply **cluster**: grouping of phrases, provided with the MemeTracker dataset. The grouping is intended to capture phrases that refer to the same external event (e.g., a specific public speech).

We further use the following definitions:

alteration: a change to a phrase such that the phrase is no longer a superstring or substring of other phrases in the cluster. Common alterations in the data were transcription errors, synonym substitutions, and misspellings.

reframing: removing words from, or adding words to, a phrase such that it becomes a superstring or substring of another phrase in the cluster. This term is intended to represent the addition to or trimming of a quotation by a user.

mutation: any **alteration** or a **reframing** of a phrase.

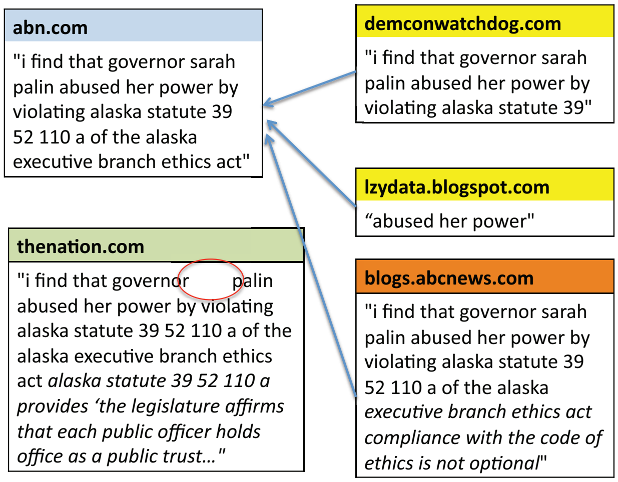

abn.com
"i find that governor sarah palin abused her power by violating alaska statute 39 52 110 a of the alaska executive branch ethics act"
demconwatchdog.com
"i find that governor sarah palin abused her power by violating alaska statute 39"
lzydata.blogspot.com
"abused her power"
thenation.com
"i find that governor palin abused her power by violating alaska statute 39 52 110 a of the alaska executive branch ethics act *alaska statute 39 52 110 a provides 'the legislature affirms that each public officer holds office as a public trust…"*
blogs.abcnews.com
"i find that governor sarah palin abused her power by violating alaska statute 39 52 110 a of the alaska *executive branch ethics act compliance with the code of ethics is not optional*"

Figure 1: An example of a quotation mutating. The original source is a report commissioned by the Alaska legislature to investigate Governor Palin's dismissal of Public Safety Commissioner Walt Monegan (Branchflower 2008). Arrows denote citation between posts.

Figure 1 shows an example of phrase mutations within a single cluster. Two copies, *demconwatchdog* and *lzydata* (highlighted in yellow), reframe the quote they are citing by copying only a subset. Another, *abcnews* (orange), reframes by adding to the quote used by the source it is citing, presumably by drawing from an additional source. Yet another, *thenation* (green), introduces a mutation not seen in other copies by omitting a word from the middle.

Filtering the Data The task of detecting mutations in memes makes sense only if one is relatively confident that the data at hand, specifically phrase clusters, represent a set of related phrases, either derived from one another, or from a single, separate source. Although the MemeTracker dataset gives us a comprehensive glimpse into what was talked about and quoted in the blogosphere during a fixed time period, a cursory examination of the clusters specified in the dataset revealed that, in some clusters, not all quotes referred back to the same source. For example, multiple people may have used similar wording, or slight variants may have been repeated on different occasions by the same person.

This can be seen by the example, shown in Table 1, where both Barak Obama and John McCain were quoted as having used the expression "lipstick on a pig". Although several distinct MT clusters contained this expression, in this case the two phrases overlapped sufficiently to be placed in the same cluster. The difference in these two phrases is not a function of the mutation of information, but rather it is the result of multiple source quotations. While these occurrences are interesting in understanding the evolution of memes over long time periods, their ambiguous provenance makes them difficult to analyze.

Table 1: An example of multiple source quotes being clustered together.

Quote	**Original Source**
You can put lipstick on a pig. It's still a pig. *You can wrap up an* *old fish in a piece of paper* *called 'change'. It's still gonna stink. . . .*	Barak Obama Sep. 9, 2008
You can put lipstick on a pig, [but] it's still a pig, *in my view.*	John McCain Feb. 1, 2007

Another source of noise is the spurious placement of a short substring into a cluster of a longer phrase containing that substring. An example can be found in a phrase cluster derived from a series of blog and news sites relaying the story of an athlete's battle with illness.

> *"It teaches you to be patient when you are lying in a hospital bed and that was almost the same strategy I chose here to wait for my chance in the pack"*

However, also included in the cluster is the phrase, *"you are lying"*. Although this is a proper substring of the previous phrase, an examination of the blogs it was extracted from confirms that this particular short quotation was used in unrelated contexts.

To determine which instances of phrases in phrase clusters were actually referring to the same event, we applied two strategies. The first was to limit our analysis to just those phrases extracted from posts which link to other posts within the same cluster. Posts that reference one another and include similar quoted text are highly likely to be derivatives of one another. The results of this filtering lend themselves naturally to network analysis techniques, and the findings on this subset of the data can be found in Section 4.2. This first filtering method resulted in a directed network composed of 9,208 nodes and 61,511 distinct edges. Although the network data was precise, this filtering technique left only a small portion of the data. The output represented 29% of the phrase clusters, 11% of the unique phrases, and 3% of total mentions present in the raw data.

The second filtering approach windowed the data to a short time period of a few days, the rationale being that mentions of lexically similar phrases that are chronologically proximate are more likely to be related. For each cluster the window was centered on the 24 hour period when the phrase cluster showed the greatest amount of activity, i.e., the "peak window". We then included all activity in the 48 hours preceding the peak window and for 48 hours after it. This approach effectively filtered out the spurious and unrelated phrases in addition to providing a way of comparing time evolution across clusters. By filtering in this manner we retain 46% of total mentions, 100% of phrase clusters and 68% of unique phrases.

Upon further manual inspection of the clusters, we found that a number of them contained short phrases that had no tie to a specific event (e.g., "I love you", "world of warcraft", and "good morning america"). A fortuitous feature of the MemeTracker clustering algorithm is that short phrases are grouped early in the clustering process and retain a low cluster ID. By discarding the first one third of the clusters, we were able to eliminate most such generic phrases. Varying the portion of cluster IDs discarded produced qualitatively similar results.

4 Patterns of Alteration: staying or straying

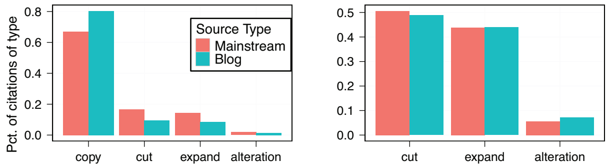
Following the filtering procedures, we were ready to examine the patterns of change resulting from meme diffusion. The data presents many examples of quotations being inadvertently altered. For example, in his acceptance speech for the democratic party nomination at the DNC convention on August 28, 2008, Barack Obama spoke the sentence

> *"John McCain likes to say that he'll follow bin Laden to the gates of Hell, but he won't even follow him to the cave where he lives."*

Interestingly, 75% of unique phrases in that cluster, and 84% of all mentions use an incorrect (e.g., "go to the cave") phrasing. The fact that so many sources published the same incorrect alteration suggests that few of them actually drew from the primary source, and are instead quoting a quote that is itself derivative. In this case, the alteration was nearly instantaneous, as it appears that one or more individuals live-blogging this event introduced the change, which was subsequently copied.

4.1 Mainstream media versus blogs

While many bloggers may have propagated the erroneous quote from Obama's speech, in that particular instance the New York Times used the correct phrasing when reporting on the event. One might ask whether mainstream media sites tend to quote with higher fidelity in general. This is an interesting question given that both news generation and consumption are increasingly shifting to social media (Purcell et al. 2010).



(a) (b)

Figure 2: (a) show the ratio of citations with different patterns of mutation. (b) shows the ratios only for those citations that introduce some form of change to the phrase.

Using the categorization of source type from the MemeTracker dataset (*mainstream* or *blog*) we checked whether one or the other was more likely to make alterations. For this we used the subset of data where we know the quote provenance as the article or blog post contained links to the source being copied from. Overall, blogs were more likely than mainstream media to simply copy exact quotes from elsewhere. However, when considering only mutations, that is changes in the quote, alterations were approximately 5% more likely to occur in pure-blog sites than in mainstream media (χ^2 = 22.55, df = 1, $p < 0.001$). The results were not sensitive to the definition of mainstream vs. blogs. They remained qualitatively similar when we reclassified some user forums hosted by mainstream media sites into the blog category. The patterns, shown in Figure 2, might be expected. Professional journalists are more likely to draw in additional material from the source, thus reframing the quote. They may also be more likely to shorten quotes for readability. However, prizing accuracy, professional journalists would be less likely to alter a quote.[*]

4.2 Citing and quoting

With the line between mainstream and social media blurring, we also wanted to examine the relationship between the prestige of a source and the fidelity with which it quoted and was quoted by others. A natural measure of prestige of a source, which is blind to its mainstream status, is the count of the number of citations a source receives. One of the advantages of using the MemeTracker dataset is that, in addition to the quoted strings, the links from each of the millions of blog posts were extracted. We used this link data to construct an inferred citation network between the different mentions within each phrase cluster.

To construct this network we ordered all mentions of any phrase within each phrase cluster by the time of appearance in the dataset. From each blog page containing a quoted phrase, we take all links pointing to other posts in the same phrase cluster. The presence of this type of link implies that the author of the post was aware of and linked to another post that contained the same or similar quote.

Thus the nodes represent different hosts (e.g., politico.com) and edges represent a link from one post to an earlier one in the same cluster. The maximum indegree (number of citations) for any node was 1,858, while the maximum out-degree (the number of times a particular node cited others) was 761. Instances of multiple edges between a pair of nodes were translated into an increased edge weight and are taken into account for measures of node centrality and prestige.

Since both the citing and cited blog post may host several phrases from different clusters we applied an additional filter to match the same quoted phrase across hosts. We only kept edges between phrases if they were identical, one phrase was a substring of the other, or if one phrase was within a certain threshold of similarity to the other. We defined this similarity threshold as a difference in length of no more than 3 characters or 3 words. Additionally the Levenshtein edit distance between the phrases could not exceed 10% of the shorter phrase's length.

We examine two questions, whether the prestige of a source correlates with others quoting it more faithfully, and whether a source that bridges more communities may spawn a wider range of derivatives than one that is nestled inside of a single community. To capture the fidelity with which sources were quoted, we recorded the difference in the number of characters between the phrase extracted from the cited

[*] Common alterations by mainstream media were often not true alterations: use of unicode characters and changes to British or American spelling.

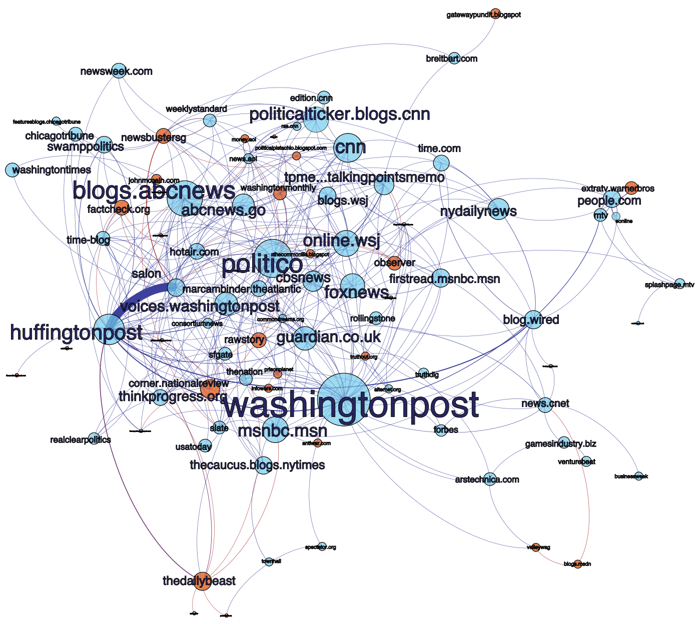


Figure 3: The most frequently cited blogs (red nodes), and mainstream news sources (blue nodes). The size of a node is proportional to the number of times that it was the source of a quotation. Nodes closer to the center have were cited more often than they cited others.)

post and the one from the citing post, and set the weight of the edge to the absolute value of that difference. Link based prestige measures we examined were in-degree, hub and authority scores via the HITS algorithm, and PageRank. In addition, we considered betweenness, which was moderately correlated with in-degree ($\rho = 0.51, p < 0.001$).

The in-degree of a source (the number of citations it received) had a positive correlation ($\rho = 0.45$, $p < 0.001$) with the percentage of citations of that source that mutated the phrase, either through reframing or alteration. This correlation was also significant, albeit weaker, for the authority, PageRank, and betweeness of a source ($0.20 < \rho < 0.34$). However, both the relative and absolute difference in length between the cited quote and the copy, representing not just the likelihood, but the amount of change that occurred, was uncorrelated with network-based centrality measures of the cited source. Interestingly, betweenness centrality of a source was moderately correlated ($\rho = 0.43$, $p < 0.001$) with the standard deviation of difference in length of the copies. This suggests that if a source straddles different communities, these communities may alter the information differently.

Finally we examine whether the authority of a source itself might give it more leeway to introduce changes. In fact, it is plausible that a source becomes an authority because it adds value by introducing changes. We find only weak support for this, in that there is a very modest correlation between the in-degree of a source and the percentage of quotes that it mutates when copying ($\rho = 0.14, p < 0.001$). This mild correlation is consistent with the slight tendency, noted above, of mainstream media sources introducing mutations at a higher rate.

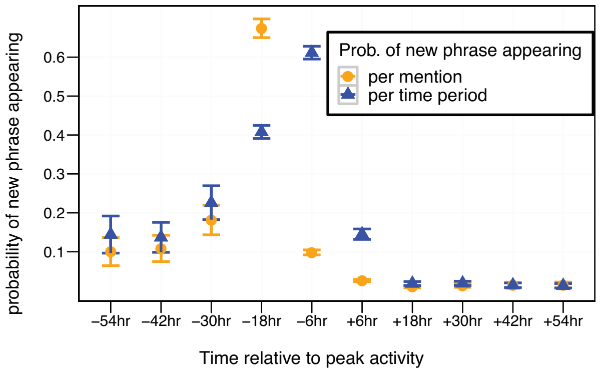


Figure 5: Rates of occurrence of new phrases, per mention (orange circles) and per time interval (blue triangles).

5 Timing and popularity of changes

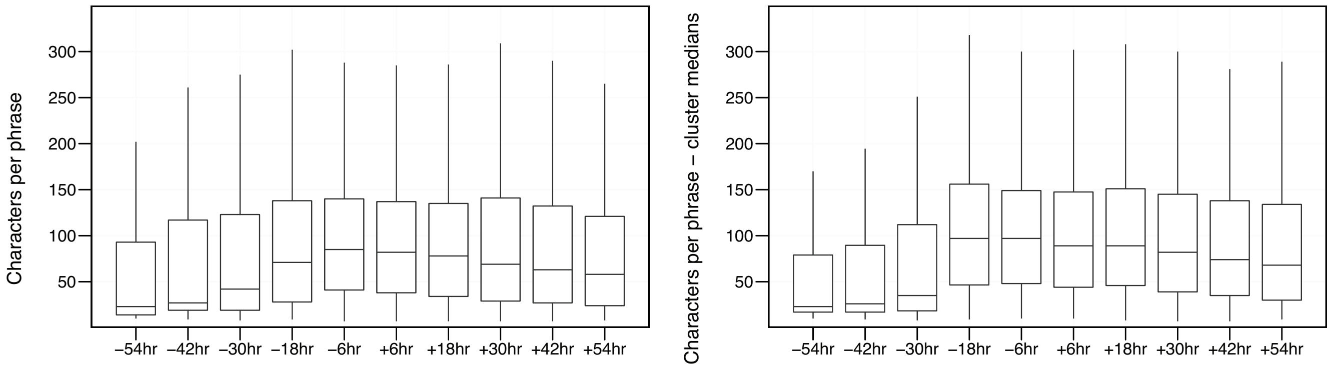
In the previous section we examined how differences in rates of change depend on the type and authority of a source. We limited ourselves to examining just quotes found in posts linking to other posts. In this section we focus on the temporal patterns of phrase variation, where a known transmission route is less important, and we can use the larger data set.

5.1 When new phrases appear

To quantify the rate novelty within a cluster, we trace the appearance of new phrases during a five hour window centered around the 24 hours of peak activity for the cluster. As shown in Figure 5, the probability that any given phrase is a new variant is highest almost a full day before the peak. However, the time when a new phrase is most likely to occur is the early period of peak activity, while phrases are proliferating. Subsequently the probability that a new variant is encountered drops, as most variants have already been created and additional mentions tend to be copies of previous ones.

5.2 Adding or chopping?

As the rate of appearance of novel phrases is increasing ahead of peak activity, so is the average length. Figure 4 shows that phrases actually tend to lengthen during that early period of a spread of a meme. Length remains constant throughout the peak, and following the peak phrases grow shorter again. A possible explanation is that during the early period, different portions of the original source are preferentially copied, with the earliest perhaps too rushed to quote extensively. Quotes that occur much after the peak are likely to be copies of copies. They are more likely to be getting shorter with successive reframings. In Section 6 we show how just such a simple copying process can replicate these patterns.



(a)

(b)

Figure 4: (a) shows the median phrase length for all phrases observed in each 12 hour period during the 5 day window of peak activity for each cluster. (b) shows the median phrase length again, but this time the data plotted is the median phrase length for each cluster in the 12 hour window. This equalizes the contribution of each cluster to the distribution of phrase length over time, but shows a very similar trend to (a).

5.3 Length and popularity

So far we've determined that phrases that appear after a phrase cluster has peaked tend to be shorter. Here we examine whether shorter phrases tend to be more popular. We expect that, just as shorter gene sequences are more likely to be preserved during evolution (Castillo-Davis et al. 2002), there may be evolutionary pressure on memes to be as simple as possible. Support for this is found in Figure 7, which shows that phrases that are shorter relative to the maximal phrase in the cluster tend to be more popular. There are two effects at work. The first is that shorter variants are actually more popular within a cluster, but also that clusters that have higher numbers of mentions have had enough sustained activity for shorter variants to have been generated. Interestingly, there is no correlation between the relative length of a phrase and how much of the 5-day interval around peak activity it spans ($\rho < 0.01$).

6 A simple copying model of meme evolution

In the previous sections we described a number of regularities in the popularity, length, and rate of arrival of new meme variants. In the original MemeTracker paper, a model incorporating recency and imitation was shown to be able to replicate the popularity dynamics at the level of the phrase cluster. However, this model does not distinguish between variants of a phrase, nor does it account for the generation of variants or resulting patterns. In this section we present a simple model that can account for the observed patterns *within* a phrase cluster.

Imitation is often modeled using Polya's urn, where at each iteration an item is selected from the urn, copied, and both the original and copy are returned to the urn. Polya's urn model, in practice, has been used to model the popularity of tags in online bookmarking systems (Golder and Huberman 2006). Crandall et al. (2008) developed a networked urn model where the probability of drawing a particular item from the urn for any given user depends on the actions of other users of the system as well as their own past actions. These prior models have assumed a fixed number of immutable choices, with a perfect copy being added to the urn following each draw. Such a copying mechanism would produce realistically skewed frequency distributions over phrases, but the popularity of individual phrase variants would stabilize over time and be independent of phrase characteristics such as length. The fixed number of immutable choices also cannot account for meme variants being created over time and from each other.

We therefore need to introduce a few simple modifications to the urn model. As in Leskovec et al., we account for recency of an item in its selection, with the probability of an item being copied decaying with the age of the item. However, since different variants are introduced through the copying process, we account for the recency of each copy separately. We also differentiate copying from the source and copying from an item that is already a copy. The parameter α allows us to tune the preference between copying from the source or from a copy, while the parameters γ and β capture the decay in probability of copying from the source or a copy as a function of time. We let t_i denote the time that copy i was created, with the source appearing at time $t = 0$. Our simulation runs as follows:

1. With probability proportional to $\alpha * e^{-\gamma * t}$, the item to be copied at time t is copied from the unique source, and not from one of the copies in the urn.
2. With probability proportional to $(1 - \alpha) * e^{-\beta * (t - t_i)}$ an item i is copied from the urn.
3. The copy is cut at two random points, with the middle portion, which can be the entire string itself, returned to the urn along with the original.

The above model simulates the process of quotes being pruned over time. The probability of returning to the source decays at a slower rate than the probability of copying one of the derivative copies, i.e., $\gamma << \beta$. This very simple model is able to reproduce both the rise and fall in phrase length over time (Figure 6), and the greater popularity of short phrases overall (Figure 7). Being so simple, it does diverge from the observed data in at least one important way. It produces many more variants, with the most popular variants taking longer to reach the popularity of the observed data. This is likely because it blindly cuts a string at any location, whereas actual quotes have a reduced number of viable cutpoints, and only a few of them select substrings of above average "fitness." Incorporating fitness as a variable will be an interesting direction for future work.

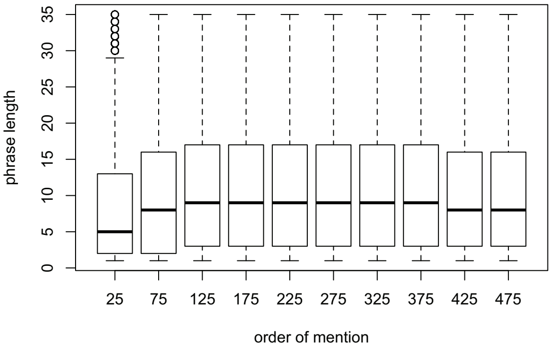
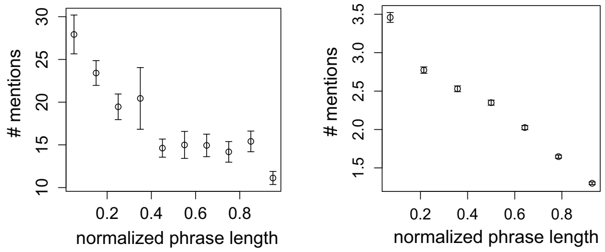


Figure 6: The evolution of phrase length produced by a simulation of a simple, imperfect copying model ($\alpha = 0.2, \beta = 0.1, \gamma = 0.005$).



(a) MemeTracker (b) Simulated

Figure 7: (a) Popularity as a function of length of phrases from the MemeTracker data. (b) simulated data from a modified Polya's urn. Phrase length is normalized by the maximal phrase length in the cluster and phrases with < 3 characters are omitted from (b).

7 Copying from a single source

So far we have shown distinct patterns in the arrival time, length, and popularity of phrase variants in the MemeTracker data, and have illustrated, using a simulation, how a simple, but imperfect, copying mechanism can produce such patterns. Specifically, this mechanism needs multiple generations of copies to be created in order to produce these patterns. However, one might argue that the patterns might arise from an alternate mechanism, one where everyone copies from the same source directly.

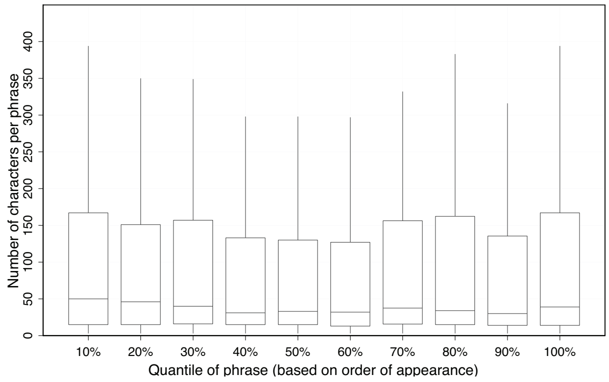


Figure 8: Blogrunner – Number of characters in each quote found on a blog citing a New York Times article, by relative order of appearance.

Consider the patterns where phrases initially lengthen, and then shorten following the peak period. Could such a pattern arise when everyone is actually copying from the same source? It may be the case that early quotes are rushed or based on only partial information. Someone liveblogging an event may not have enough time to copy a long quote. Or they may have been given only parts of a document (e.g., Obama's prepared speech) ahead of the event. During and shortly after the event, but still likely before the peak of activity, the individuals who are writing the quotes may be most interested in creating unique quotes that would set them apart and appeal to their audience. They might produce the lengthiest quotes. As time goes on, quotes are copied by less invested individuals who may be content to make exact copies or to abbreviate.

In order to test whether this could be the case, we examine the BlogRunner dataset described in Section 3.2. As mentioned, each quotation is extracted from a post citing a New York Times article. First, we check that the blogs are not citing one another, which would increase the likelihood that they are copying from secondary sources. Since only a small portion of posts (4.8%) do so, we remove them from the analysis. We then examine the relationship between when a quotation appears and its length. Figure 8 shows a decreasing trend in phrase length over time, less pronounced than in the MT data, and no initial increase in length. This suggests that while some of the decrease in length of phrases over time may be attributable to fatigue while copying from a single source, most of the observed pattern is consistent with multiple degrees of propagation, as simulated in our model.

8 Conclusions and Future Work

In this paper we presented the first large-scale quantitative study of meme mutation in social media. Our data was limited to quoted text derived from two sources, the MemeTracker dataset and blogs citing New York Times articles. Since quoting activity implies copying information with high fidelity, one would expect minimal changes in content. Yet we find that changes are very common.

Interestingly, changes are more likely to be introduced by mainstream media sources. Typically these changes are abbreviations or expansions of source quotes, which might add value to readers of these outlets. On the other hand, blogs are more likely to simply copy. When they do introduce changes, they are more likely to be alterations, which may be unintentional. More frequently cited sources, whether blog or mainstream, and sources that bridge different communities, are quoted more variably. These findings have important implications for the fidelity of information that is being consumed online, especially since internet users are increasingly accessing new content via social media.

We further find temporal and structural patterns that arise from this transmission process: shorter phrases are more likely to persist, and average phrase length peaks around the time of peak activity. The persistence of short phrases could be due to several factors. First, a short phrase is less vulnerable to alteration, since it has fewer potential cutpoints and mutation points. It may also be easier to copy exactly. The peak in phrase length can be explained by a compound effect between the (longer) source being more likely to be copied initially, and the gradual shortening of phrases with each successive copy. We have captured these empirical findings and intuition in a simple model of imperfect copying.

In future work we would like to apply linguistic analysis to detect attributes of highly propagated and highly cited phrases. Similarly, we plan to modify our simple model to assign different levels of fitness, and hence different likelihood of being preserved, to different parts of any given text. We may further model a networked urn, where the probability of drawing a particular phrase depends on the draws of other nodes one is connected to via social media.

Finally, our analysis so far has not taken into account any of the text surrounding the quotes. An examination of this text would allow us to determine whether there is a connection between mutation and sentiment, e.g., if strong sentiment is more likely to accompany a mutation than a perfect copy. We could also take into consideration the community structure of the networks through which the information is propagating, to see whether information is more likely to mutate within or across community boundaries.

Acknowledgements This work was supported in part by NSF IIS-0746646 and MURI award FA9550-08-1-0265 from the Air Force Office of Scientific Research.

References

Adar, E., and Adamic, L. 2005. Tracking information epidemics in blogspace. In *Web Intelligence'05*, 207–214.

Balahur, A.; Steinberger, R.; Goot, E.; Pouliquen, B.; and Kabadjov, M. 2009. Opinion mining on newspaper quotations. In *WI-IAT'09*, volume 3, 523–526.

Barzilay, R., and Lee, L. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HTL/NAACL*, volume 1, 16–23.

Berendt, B., and Subasic, I. 2009. Stories in time: A graph-based interface for news tracking and discovery. In *WI-IAT '09*, 531–534.

Branchflower, S. 2008. Report to the Alaskan legislative council, Oct. 10, 2008.

Castillo-Davis, C.; Mekhedov, S.; Hartl, D.; Koonin, E.; and Kondrashov, F. 2002. Selection for short introns in highly expressed genes. *Nature Genetics* 31(4):415–418.

Copus, M. 2011. Chinese whispers on twitter sparks panic in london. http://mediadigest.co.uk/news/chinese-whispers-on-twitter-sparks-panic-in-london Retrieved 02/02/2011.

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J.; and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *KDD '08*, 160–168.

Ennals, R.; Trushkowsky, B.; and Agosta, J. M. 2010. Highlighting disputed claims on the web. In *WWW '10*, 341–350.

Fisher, D.; Hoff, A.; Robertson, G.; and Hurst, M. 2008. Narratives: A visualization to track narrative events as they develop. In *VAST '08*, 115 –122.

Golder, S., and Huberman, B. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* 32(2):198–208.

Gomez Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. In *KDD'10*, 1019–1028.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW '04*, 491–501.

Hall, D.; Jurafsky, D.; and Manning, C. D. 2008. Studying the history of ideas using topic models. In *EMNLP '08*, 363–371.

Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2005. On the bursty evolution of blogspace. *WWW* 8:159–178.

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2007. The dynamics of viral marketing. *ACM Trans. Web* 1.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD'09*, 497–506.

Mei, Q., and Zhai, C. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05*, 198–207.

Nallapati, R.; Feng, A.; Peng, F.; and Allan, J. 2004. Event threading within news topics. In *CIKM '04*, 446–453.

Purcell, K.; Rainie, L.; Mitchell, A.; Rosenstiel, T.; and Olmstead, K. 2010. Understanding the participatory news consumer. *Pew Internet and American Life Project*.

Shinyama, Y.; Sekine, S.; and Sudo, K. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of Human Language Technology Conference*, 40–46.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *WSDM '11*.