Computational Treatise on Urban Legends

Coen D. Needell

Introduction

Urban Legends, or Contemporary Legends, are a type of folklore popularized by Jan Harold Brunvand in his book: *The Vanishing Hitchhiker: American Urban Legends & Their Meanings*. Urban Legends are considered to be the modern continuation of the human tradition of folklore and legends. Folklore does not occur exclusively in so-called primitive or traditional societies, and by studying modern folklore in the same way that we study the folklore of traditional societies, we can also learn about modern societies.(Brunvand 2003) Urban Legends, like their past counterparts, form as complete stories with plot and characters, although the exact identities of these characters tend to stay vague, unlike mythic stories.

We have to focus our attention though and draw boundaries between legend, fake news, conspiracy theory, conjecture, and so on. We are all fully aware that the systems that make the Internet a breeding ground for urban legend, also make it a fantastic substrate for the other types of rumor and hearsay (Blank and McNeille 2018). For our purposes, we will consider the genre of "Urban Legend" to be defined by the fact that the people who spread the legend know that it isn't true. Conspiracy theories tend to be spread by the true believers, and fake news tends not to be spread by people, but by robots and organizations. In addition, the Urban Legend tends to take place in a fuzzy time-frame, the near past. Generally, the Legend happened in a less-specific time, the storyteller will not give a date, nor even a month in which the story occurred. Only in the most specific of legends will the story include, at most, a season in which the story occurred, but rarely a specific year. Some examples include the story of the arcade cabinet, Polybius, that supposedly appeared in a Portland area bar in the summer of 1983. The story goes that the cabinet was responsible for children and teens forming a line out the door, until it eventually hospitalized a kid who had played it for too long. Many legends will give you a decade, or an era. The story of the Foo Fighters, mysterious lights in the sky, took place during World War II, fueling speculation and conspiracy theories about the secret technology of the Third Reich. There are real government documents that report sightings of Foo Fighters during the late 1930s and early 1940s, but many of the stories that circulate are apocryphal. Generally, the reports talk about seeing strange glimmers in the distance, whereas the stories you might hear on the internet talk about seeing faces emerging from lights, clouds parting, the weather suddenly changing, and pilots engaging in dogfights with ghostly adversaries.

These stories that are anchored in a time and place become unstuck in time when they're reproduced and refactored into urban legends, but most of the legends are timeless. Generally the story will take place in the near past, some time in the past two years from the time of storytelling. For example, the story of the Vanishing Hitchhiker could have occurred in any time or place, hitchiking has been a common occurrance in any society that has access to roads and wheeled vehicles. In it's modern form, the legend has been traced back to the 1870s and similar stories can be found in societies around the world.(Brunvand 2003) In general, the legend portrays a hero and a stranger. The hero picks up the stranger by the side of the road, and gives them a ride. The stranger is either silent, or makes odd conversation. Then, when the hero finally starts to become agitated by the stranger's behavior, the stranger vanishes from the moving vehicle. The vague nature of the urban legend also extends to its characters. In the story of the Vanishing Hitchhiker, the characters could be anyone. Another example is the story of the Black-Eyed Children. According to the story, a child with solid black eyes appears at the door of the hero. The child asks the hero if they can use the bathroom. When asked about their background,

the child dodges the question. The child speaks in a monotone, and has perfectly white teeth, the stories always mention the whiteness of the teeth. They speak with the diction of adults. The child continues to insist that the hero invites them in, until eventually the hero breaks down and flees, in some cases they might even call the police, only to find that there is no trace of the child as soon as they check their entryway again.

Because the characters can truly be anyone, the the stories plausibility is never tied to the identity of the character, and more importantly, the story can be spread by anyone. The storyteller generally doesn't go so far as to say that they are the hero, but it's often someone they know, but not a core friend or immediate family. This is so common that modern urban legends are sometimes called FOAF or friend-of-a-friend stories. By disconnecting the story from the storyteller, it's possible for anyone to tell the story, and still have the tension be heightened by an intrinsic familiarity with the storyteller.

Often it can be hard to tell the difference between a conspiracy theory and an urban legend. Conspiracy theories either exist in a sharp time frame, like the Kennedy Assassination, or they exist in the present and are presented as ongoing. This is a major delimiter. The conspiracy theory can take place over a variable length of time, for example stories about the Illuminati claim the organization date to anywhere from the 1800s to ancient times. The Urban Legend, however, always ends when the story ends. The hero does not adopt a black-eyed child, the driver does not get a drink with the vanishing hitchhiker every Friday going forward. Fake News on the other hand always exists in the Now, and often includes a Call-To-Action which forces it to be mixed up in current events. Fake News usually falls into the wider category of Hoaxes, being stories that are written and disseminated with the express purpose of deceiving the public. In addition, the star characters in Fake News and Conspiracy Theory are celebrities, senators, presidents. (Frank 2011) Most conspiracy theories also place the conspiracy at the foot of a specific person or group, for example the military, a president, or Woody Harrellson's father. By contrast, the identity of a character in an urban legend is always an individual, but their exact identity is vague. Urban Legend stars the everyman, the friend of a friend, the weird uncle, but it also stars some hard-to-understand entity, like a fairy, an alien, slenderman, or a large company's incomprehensible legal policies.

Another delimiter between Urban Legends and other forms of modern lore is the argumentative role of the story. The story in an urban legend fits into the same sociological niche as the folklore of old. The story is often a morality play, often teaching the audience caution. The hero often survives to tell the tale precisely because they were suspicious of the circumstances as soon as they appeared.

Some previous work on urban legends has defined them as being related to rumors. (Dunn and Allen 2005) This stance usually maintains that both rumors and legends are told with the intention of being believed, told as if they are true, and are difficult to verify. The distinction between the two being that rumors are generally smaller, simpler stories, and are not told in a narrative fashion. Consider a rumor like "I heard that John in HR is secretly dating Cassie in accounting" or even something with larger stakes like "Justin Trudeau is having an affair with Angela Merkel". These stories don't contain a narrative arc in the same way that the quintessential urban legends do.

Generally "Urban" or "Contemporary Legends" describes stories that few people truly believe, but are still told and retold time and time again. These stories often have common features, they leverage the unintuitiveness of modern life, a sense of distant credibility ("This happened to a friend of a friend"), and in some cases a sociopolitical call-to-action. Prototypical Urban Legends include the story about a man meeting a woman in a hotel bar and waking up the next day missing a kidney in a bathtub full of ice, or the story where your uncle's friend went down into the sewers and encountered a bask of exotic crocodiles that had been flushed down by irresponsible pet owners.

But let's focus for a moment on that third millennium (and also the decade leading up to it). The human experiment (although mostly the western countries and Japan), had just been linked to one another through the Internet. Before Google created the "clearnet", before Facebook acquired every social forum, we had specified forums, on separate sites, operated by separate

people. We had Usenet, a social media system built out of email and simple text-based servers.

Usenet was a system of email mailing lists that were connected and chained together in a manner such that anyone could post on them, and anyone could read the history. They functioned somewhere between bulletin board systems and email newsletters, and filled the niche that would later be overrun by forum software. The main difference being that usenet was decentralized. Usenet hosts would pass messages along to people, and receive messages from posters, often through a chain system, to ensure that a post which was sent to more than one newsgroup got to all of the requested recipients. (**???**)

Usenet was divided into newsgroups, which were in turn divided into subgroups. Generally these would be addressed by their hierarchy, separated by a period. For example, `sci.agriculture.beekeeping` would be a board about beekeeping, but it also fit under the subcategory agriculture, which was in the science hierarchy. There were a number of hierarchies which were controlled by specific groups. The major groups are referred to as the Big 8. Originally, in 1980 when Usenet was created there was only one major newsgroup, which was called `net.*`. In 1987, Usenet went through a major restructuring, creating the Big 7. These were intended to be an encompassing set of categories, and they were: `comp.*`, for discussion of computer related topics, `news.*`, for discussion of Usenet itself, `rec*`, designed for recreational activities like games and hobbies, `sci*`, discussing science, `soc.*`, for both socializing and for discussing social issues, `talk.*` which was targeted at discussion of contentious issues like religion and politics, and finally `misc.*`, things that don't fit in the other Big 7. This restructuring led to a controversy about what was allowed on usenet. Even `misc.*` was heavily moderated. The rules for each hierarchy were set up as to prohibit, from any of the big 7, discussions about recreational drug use, sex, or sharing recipes. So the self-styled Usenet Cabal set up an eighth major newsgroup, `alt.*`. Designed to keep the debauchery of the internet contained. `alt.*` is not considered an official member of the Big 8, and the real eighth member was added in the 90s, `humanities.*` for discussion of the humanities, in a similar spirit to `sci.*`. The `alt.*` newsgroup however, became a free-flowing, unmoderated message board, with all of the issues that includes. The Usenet community would often jokingly refer to `alt.*` users as "Anarchists, Lunatics, and Terrorists."

In a time before Google, or even Snopes, you couldn't fact check that troll's rant, and those rants spread like wildfire.(Dunn and Allen 2005) (Donovan 2004) Even in the later times of transition, we saw legends like the story of slenderman, video game related legends like "Herobrine", and so on. Obviously we see the scarier legends show up on forums dedicated to horror stories and urban legends, we see video game legends sprout up out of those games' communities, and we see less intense legends appear in more general use communities(Blank 2007). These are things like the story about the Neiman Marcus cookie recipe. This brings us to the question: How do Urban Legends spread through online communities, and how do those communities effect the nature of the legend itself?

Methods

Data Sources

Much of the early internet is made available to us. Groups like the Internet Archive host repositories full of early internet interactions. Many of the forums that people used to discuss broad topics are still existing, and keep their old posts up for posterity. That being said, this is far from complete, IRC (Internet Relay Chat) logs will be mostly wiped from the record. IRC servers were often private, and hosted by private entities. In addition, the Usenet archive that will be used for this discussion is not complete, it is however very comprehensive for the boards that it exists for, and this includes the largest boards. By compiling usenet records, we see a very large section of the picture. We will be able to see links between individuals, and links between communities (archive.org 2020). I will focus on a couple of usenet newsgroups, `talk.*`, `soc.*`, `rec.*`, and `sci.*`. I have also procured data from select `alt.*` newsgroups, including `alt.folklore.*`.

Using text classification methods, we can sift through these posts, and pick out a selection of urban legends that we will use to study their dynamics overall. Then, we can find posts that reference those legends. Conceptually, we can

imagine that posts fit into one of three categories, irrelevant, which consists of posts that do not reference the urban legend at all, carrier, which consists of posts that reference the story in passing, but don't tell the story, and spreader, which consists of posts that tell the legend, thus spreading it to other people in the community. We can imagine that a single user can write posts that belong to all of these categories, but each individual post can only belong to one.

The spreader posts will be generally longer, and will likely not differ much from one another, the carrier posts, however, will rarely be similar. By way of explaining this, consider the case of the Foo Fighters. Since they only reference the phenomenon, only a small portion of their verbiage is devoted to the story. As such, you can imagine a carrier post that references Foo Fighters but is really all about the nature of paranoia in conflict, and you can imagine another carrier post that uses Foo Fighters as way to springboard into a discussion about the government hiding secret technology from the public. Both of these posts reference the story, and both of them do the same job, serving as a vector of initial exposure to the story, but text classification systems and topic modeling would not associate them unsupervised. Compare this to two spreader posts. Since they both serve to tell the story, the bodies of the post would have to be mostly devoted to telling the story. Even if one was leveraged at telling the story of the Foo Fighters to aviation experts, and therefore might employ aviation jargon to make the story more believable, and another was leveraged toward a more general audience, the bulk of the story is the same. Because of this, a text classification system or topic modeling would label these two posts as more similar.

Data Modeling

Since these data sources size in the hundreds of gigabytes of unstructured text data, they need to be organized in such a way to make analysis easier both on the scientist and on the computer. In their raw formats, usenet posts are stored in a mailbox file. Posts are delimited by a line that reads `FROM` and then a long string of numbers. They then have a block of metadata, followed by the body. Usenet allows many optional metadata fields, but requires a few.

The data model reads through a mailbox file, and extracts the only the required fields of metadata, storing it as attributes of the `Post` object. The `Post`s are sorted together into `Newsgroups` which are sets that are inherently tied to a zip file on the computer. They have special methods to save and update the `Newsgroup` from disk. They can be constructed either by providing a list of posts to add to the `Newsgroup` and a file name for the zip file, or using an alternate constructor. One of these can construct a `Newsgroup` by specifying a `.mbox.zip` file as provided by archive.org. This will take a while. Another can do this by specifying a newsgroup as specified by the Usenet protocol. This will take longer.

The `Newsgroup` acts like a normal set, but it also has methods to save the current state of the set to the zip file, and to load a zip file constructed by a `Newsgroup`, adding those `Posts` to it.

`Posts` have a unique identifier, which is delimited by the Message-ID: metadata tag. Since these are unique for all of Usenet, they can be used to tell if two posts are identical across newsgroups. The `Newsgroup` class is implemented such that it will detect this and not repeat two `Posts` with the same message ID. `Newsgroups` can then be iterated over like any other set, and extracting information from them is reduced to an $\mathcal{O}(1)$ operation.

For more information see the `analysis.py` file stored at https://github.com/UC-MACSS/final-research-paper-SoyBison.

Analysis

Using those three categories above, we would construct a temporal model of spread through the network that will let us understand how the legend spreads overall. This is a contagion model of ideas that is based off of the idea that the individuals in the community are exposed to a legend, and then spread it for some amount of time, before they get bored of it and stop. In addition, we can use topic modeling to create a set of features for all of the spreader posts for all of the urban legends that we're studying, and see how the spreader posts change their telling of the story based on the community that they're talking to. For example, we'd expect the story of slenderman to take on a different tone in an

anthropology focused community than it would in a parenting focused community.

The contagion model posits that each individual in the community is either *infected*, *susceptible*, or *recovered/dead*. These are sometimes called SIRD models. Since an Urban Legend (mostly) cannot kill, this model will not consider anyone to be dead.

Ideally the topics that we deconstruct from the spreader post serve as analogues for the "genes" in the Darwinian model of idea spread (Dawkins 1989) (Kronfeldner 2014). Under this theory, we would expect that the topics in a spreader post will change as the legend interacts with a new substrate, a specific type of community **(Citation Needed)**. Previous work on the Darwinian model of ideas focused on the concept of emotional selection. Ideas that employ emotional techniques, pathos argumentation, or evocative imagery tend to spread faster than those that don't.

We would also expect that some stories are less prone to mutation than others. For example, there is a certain subgenre of scary story that appeared on the early internet called "creepy pasta" which is a mutation of "copy pasta" which is in turn a mutation of "copy paste". These are stories which mostly appeared in the same format, implying that a person spreading the story would be copying the story from the place that they first saw it, and pasting it to a new post. While "Creepy Pasta" have died down over the years, and now the term is synonymous with internet scary stories, "copy pasta" is still a core part of how internet communities interact, with many of the famous copy pasta becoming memes of their own, often mutating in such a way that they change topics entirely, but keep the cadence and tone of the original. (Heath, Bell, and Sternberg 2001)

If the Darwinian model is accurate, we should expect the topics that make up an Urban Legend to change when the story is introduced to a new environment. Ideally the topics that change will be in some way connected to the new substrate's nature. For example we would expect the version of an urban legend on an academic community to make more reference to complex ideas and theory, whereas in it's original substrate, the language would likely be more general.

Results

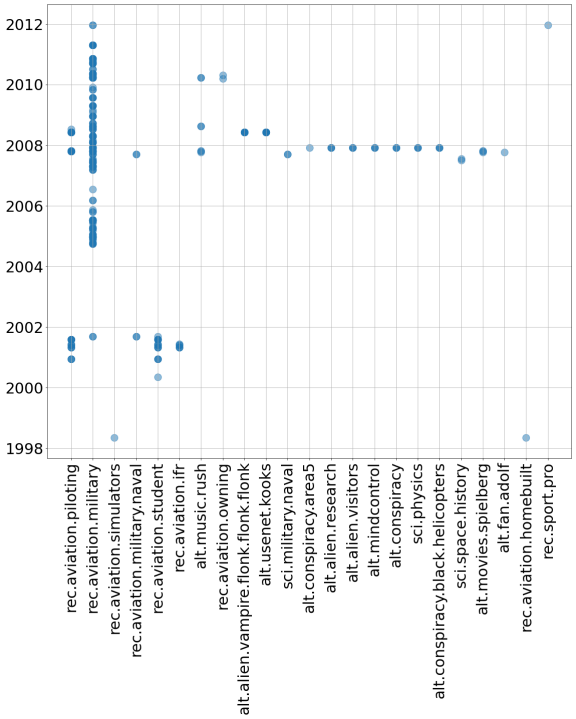

Figure 1: Posts from `rec.aviation.*` organized by date posted, and shown with corresponding cross-post newsgroups.

I have done some preliminary analysis of the data. As a sample population, I have looked at occurrences of the story of the Foo Fighters. For this case, I looked only at posts that were sent to `rec.aviation.*` newsgroups. Since a single post can also be cross-posted to other communities, we can see where else these posts ended up, and how that location changed over time. First, I'd like to point out that there are two clear clusters in time that the foo fighters story is mentioned outside of military and piloting communities. They are mostly cross-posted to other newsgroups about the military, or to groups that are about aliens, conspiracy, and the paranatural.

We see a sort of trumpet-shape, with posts referencing foo fighters starting out as being only posted to aviation themed newsgroups. Later though, these posts are being cross-posted to message boards about the paranatural. In addition we see that clusters of foo fighters related posts within the military community surge and then trail off in time.

There are also small clusters of posts during the earlier surge in the `rec.aviation.student` and `rec.aviation.piloting` communities. These are communities that are focused on piloting techniques and new pilots. Unlike in the military community, we see a slower pickup, but a faster trail off in infections.

Discussion

The general shape of how these stories spread is consistent with a contagion model of idea spread. The story has as it's hallmarks a pilot, who is flying a plane at the time of contact with a strange entity. It makes sense that these stories would start out in message boards aimed at pilots. In addition, the story is sourced from military personell, so we expect the story to do well in a substrate made up of military-related individuals, which is consistent with the observations. 83% of posts that reference foo fighters were posted to the `rec.aviation.military` newsgroup.

Within the military community, we see that clusters of foo fighters related posts surge and then trail off. This is consistent with our model as well, for most of the military personell, this story is more common and familiar, so although modeled infections happen quickly, the spreading doesn't continue on for long after that, as the number of resistent nodes in the network rises.

Consider now the `student` and `piloting` newsgroups. These are focused on newcomers to the community and discussion of techniques. Ostensibly, these would be worse substrates for spread, so we see slow build-ups, and then the community gets bored with the story faster, so posts drop off sharply.

Conclusion

Although only preliminary, these results are still exciting. The time-series data about foo fighters related posts seem to follow the contagion model of idea spread fairly well. At the moment, though, I cannot say anything about the content of the posts, whether they are spreading the stories as true or as false, if they contain aviation jargon, etc. This will make for more study as this project goes on. For now though, the results are promising.

In the future I propose analyzing the same legend across more hierarchies, to get a better idea of posts that find themselves primarily outside of the origin newsgroup. Although quite informative, this preliminary analysis cannot tell us anything about how the story spreads through other communities, and that would stand as an important point in the theory.

References

archive.org. 2020. "Internet Archive: Digital Library of Free & Borrowable Books, Movies, Music & Wayback Machine." https://archive.org/.

Blank, Trevor J. 2007. "Examining the Transmission of Urban Legends: Making the Case for Folkore Fieldwork on the Internet."

Blank, Trevor, and S McNeille Lynn. 2018. *Slender Man Is Coming: Creepypasta and Contemporary Legends on the Internet*. Utah State University Press.

Brunvand, J.H. 2003. *The Vanishing Hitchhiker: American Urban Legends and Their Meanings*. W. W. Norton. https://books.google.com/books?id=eY-W/_LiKe18C.

Dawkins, Richard. 1989. *The Selfish Gene*. New ed. Oxford University Press.

Donovan, Pamela. 2004. *No Way of Knowing: Crime, Urban Legends and the Internet*. Routledge.

Dunn, Henry B, and Charlotte A Allen. 2005. "Rumors, Urban Legends and Internet Hoaxes." In *Proceedings of the Annual Meeting of the Association of Collegiate Marketing Educators*, 85.

Frank, Russell. 2011. *Newslore: Contemporary Folklore on the Internet*. University Press of Mississippi. https://doi.org/10.14325/mississippi/9781604739282.001.0001.

Heath, Chip, Chris Bell, and Emily Sternberg. 2001. "Emotional Selection in Memes: The Case of Urban Legends." *Journal of Personality and Social Psychology* 81 (6): 1028–41. https://doi.org/10.1037/0022-3514.81.6.1028.

Kronfeldner, Maria E. 2014. *Darwinian Creativity and Memetics*. Routledge.