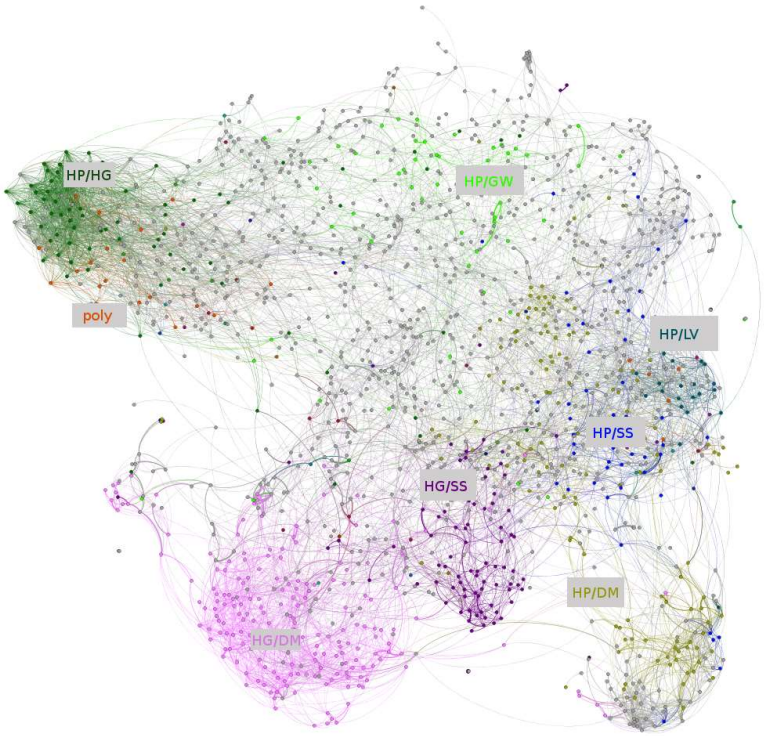
# Fanfiction, Graphs, and PageRank

*Posted on July 6, 2014*

*math (../../posts/tags/math.html), fanfiction (../../posts/tags/fanfiction.html), graphs (../../posts/tags/graphs.html), visualization (../../posts/tags/visualization.html)*

On a website called fanfiction.net, users write millions of stories about their favorite stories. They have diverse opinions about them. They love some stories, and hate others. The opinions are noisy, and it's hard to see the big picture.

With tools from mathematics and some helpful software, however, we can visualize the underlying structure.




Graph of Harry Potter Fanfiction, colored by ship

In the following post, we will visualize the Harry Potter, Naruto and Twilight fandoms on fanfiction.net. We will also use Google's PageRank algorithm to rank stories, and perform collaborative filtering to make story recommendations to top fanfiction.net users.

If you're not interested in the details, you can skip to the following:

**Interactive Graphs**: Harry Potter (graphs/HP-ship/), Naruto (graphs/NAR-ship/), Twilight (graphs/TWI-ship/)

**Story Rankings**: Harry Potter (pagerank/hp.html), Naruto (pagerank/naruto.html), Twilight (pagerank/twi.html)

**Story Recommendations**: Harry Potter (recs/hp.html), Naruto (recs/nar.html), Twilight (recs/twi.html)

And of course, you might skim below to see the pretty pictures!

## Introduction

Fanfiction is a wide-spread phenomenon where fans of different works write derivative stories. This ranges from young children writing their first stories about their favorite fictional characters, to professional-quality stories written by aspiring novelists. Many such stories are posted to websites where they are read by a large audience and commented on. The largest such website is fanficiton.net (https://www.fanfiction.net/).

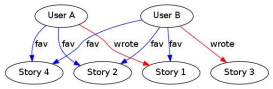
The sheer amount of fanfiction out there is rather staggering. The total number of stories on fanfiction.net exceeds six million. Harry Potter stories account for around 14% of these, followed by Naruto (around 7%) and Twilight (around 4%) (FFN Research (http://ffnresearch.blogspot.com/2010/07/fanfictionnet-story-totals.html)). The majority of these stories have very little in the way of readership, but popular stories can have a large number of readers.

Some research was done into the demographics of fanfiction.net users and other topics by FFN Research (http://ffnresearch.blogspot.com/). They found that 78% of fanfiction.net authors who joined in 2010 identified as female. Further, around 80% of users who report their age are between 13 and 17.

A lot of other interesting research and analysis has been done on the blogs Destination: Toast! (http://destinationtoast.tumblr.com/stats) and TOASTYSTATS (http://toastystats.tumblr.com/).

## Basic Methods

In addition to allowing users to post stories they write, fanfiction.net allows authors to "favorite" stories they like. Looking at which stories tend to be favorited by the same users gives us a way to understand connections between stories.




In order to analyze this, we must collect a large amount of metadata from fanfiction.net ("scraping"). We note that we don't actually collect any significant content, just a lot of data about relationships between pieces of content. Fanfiction.net's terms of service, as the author understands them, allow this with some restrictions:

> *4(E) You agree not to use or launch any automated system, including without limitation, "robots," "spiders," or "offline readers," that accesses the Website in a manner that sends more request messages to the FanFiction.Net servers in a given period of time than a human can reasonably produce in the same period by using a conventional on-line web browser. Notwithstanding the foregoing, FanFiction.Net grants the operators of public search engines permission to use spiders to copy materials from the site for the sole purpose of and solely to the extent necessary for creating publicly available searchable indices of the materials, but not caches or archives of such materials...*

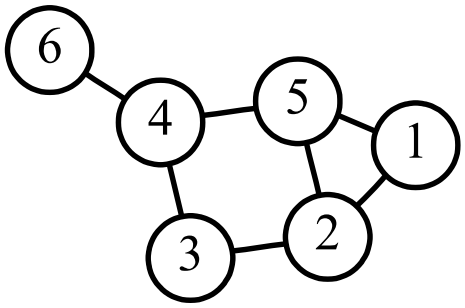
In order to ensure compliance with these terms, the author intentionally built significant rate limiting into the scraper and took care to minimize the load put on fanfiction.net. While the issue of academic analysis was not mentioned, it was not excluded and fanfiction.net's operators have not previously objected to similar academic work. Further, this work could be the preliminary research needed for someone to build a good fanficiton search engine.

Another section of the terms of service prohibits collecting personally identifiable information, which they define to include usernames. As such, I have deliberately discarded all such information and don't use it. (Though, I note that several search engines do – try searching for an authors name on any major search engine.) I do refer to some usernames in this post, but that was done entirely by hand.

In collecting data, since we are only looking at a subset of users, it is important to be wary of sampling bias. For example, if we sampled authors starting from the favorites of a particular author, or from those who had contributed stories to a community, we might get a very skewed perspective of the stories on fanfiction.net. The author considered a number of approaches, but concluded the fairest approach would be to use the authors of the most reviewed stories on fanfiction.net. This is a bias, but it should bias us towards the most interesting and important parts of the graph.

### Graph Construction

A graph (http://en.wikipedia.org/wiki/Graph_(mathematics)), in the context of mathematics, is a collection of objects called vertices joined by connections called edges. For example, cities can be thought of as the vertices a graph connected by different highways and roads (the edges).




An example of a graph (from Wikipedia)

A weighted graph is a graph where some edges are "stronger" than others. For example, some cities are connected by giant 6-lane highways, while others are connected by gravel roads. Larger weights represent stronger connections and smaller weights represent weaker ones. A weight of zero is the same thing as having no connection at all.

We will be interpreting fanfiction as a weighted graph, where edges represent a "connection" between stories. We will be using as our weights for edges the probability that someone will like both stories, given that they like one. That is, $W_{a,b} = \frac{|F_a \cap F_b|}{|F_a \cup F_b|}$ where $F_s$ is the users who favorited the story $s$.

There are lots of other possibilities, some resulting in directed graphs:

- (directed) The probability that someone who favorites $a$ will favorite $b$: $W_{a \to b} = \frac{|F_a \cap F_b|}{|F_a|}$
- The probability that someone who favorites $a$ favorites $b$ times the probability that someone who favorites $b$ favorites $a$: $W_{a,b} = \frac{|F_a \cap F_b|^2}{|F_a| * |F_b|}$
- The lesser of the probability that someone who favorites $a$ favorites $b$ and the probability that someone who favorites $b$ favorites $a$: $W_{a,b} = \min\left(\frac{|F_a \cap F_b|}{|F_a|}, \frac{|F_a \cap F_b|}{|F_b|}\right)$

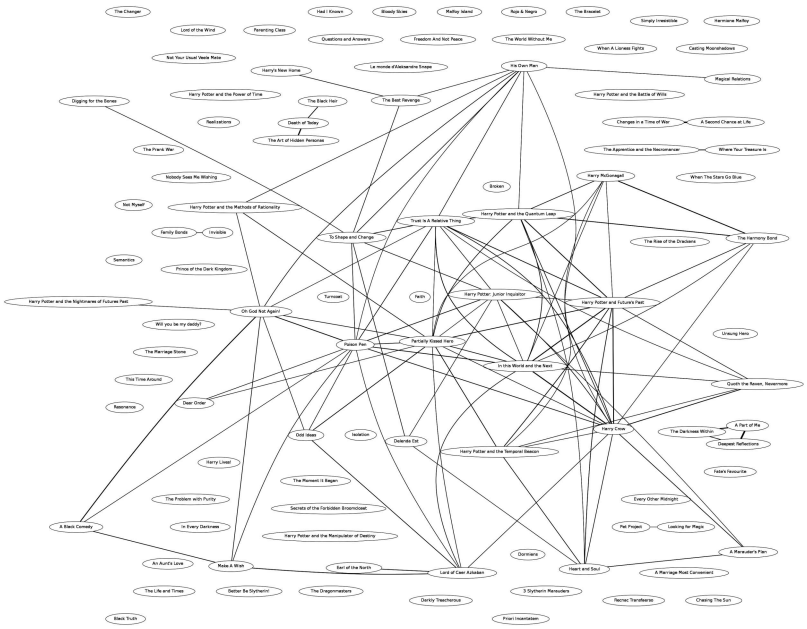
Our experience was that it didn't matter too much for the results, for large graphs.

(It's worth noting that many of these could easily generalize to higher-dimensional edges for a weighted hyper-graph.)

In our selected weight definition, $W_{a,b} = \frac{|F_a \cap F_b|}{|F_a \cup F_b|}$, we give equal weight to the preferences of all users. But there's a lot of variance between users: some favorite everything under the sun, while others very selectively favorite stories they really like. If we give the users who favorite thousands of stories the same weight as users who favorite ten, the users who favorite thousands dominate everything (and aren't a very good signal).

Instead, we give each user $u$ a weight of $\frac{1}{20+n(u)}$ where $n(u)$ denotes the number of stories $u$ has favorited. This results in a measure on the space of users, $\mu(S) = \sum_{u \in S} \frac{1}{20+n(u)}$, and the equation for our weights becomes $W_{a,b} = \frac{\mu(F_a \cap F_b)}{\mu(F_a \cup F_b)}$.

Applying these techniques to a couple of the top Harry Potter stories, we get the following graph (using graphviz (http://www.graphviz.org/)):




Small labeled graph of top Harry Potter stories

With a small amount of investigation, it's easy to understand a lot of the graph's structure. For example, on the lower right hand side, there's a triangular clique.

A quick Google search reveals that this triangular clique consists of the "Dark Prince Trilogy" by Kurinoone. The stories are more strongly linked to their immediate predecessor/successor than the pair separated by a story are to eachother.

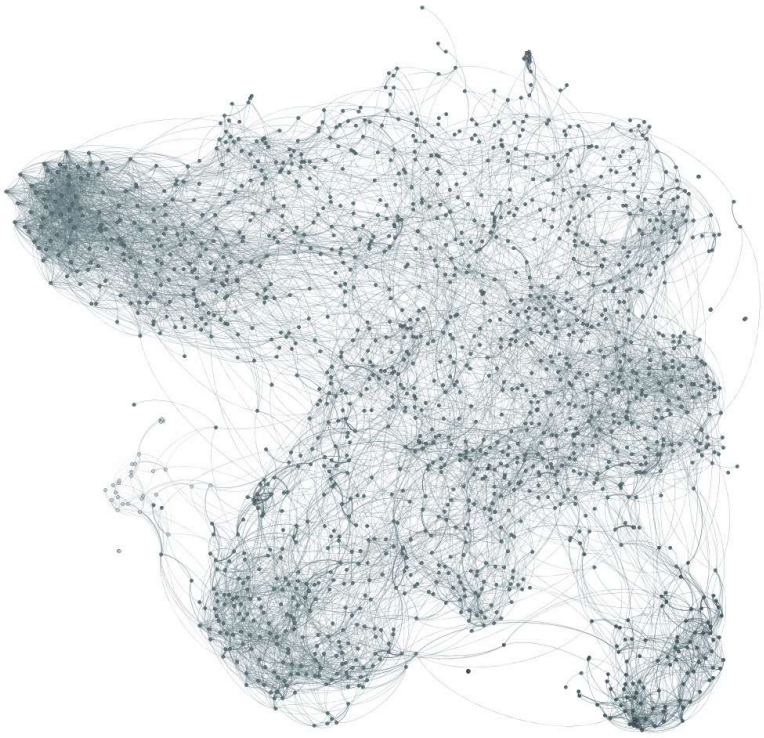

### Large Graph visualizations for Harry Potter

If we use different tools, we can visualize much larger graphs.

We consider the top 2,000 most reviewed Harry Potter stories and their authors. Based on the author's favorite lists, we construct a weighted graph, with the stories as nodes (edge weights are calculated as above).

We then prune the graph's edges, keeping the top 8,000 most strongly weighted edges. We also prune the nodes, keeping only those with at least one edge. This leaves us with a graph of 1,623 nodes and 8,000 edges.

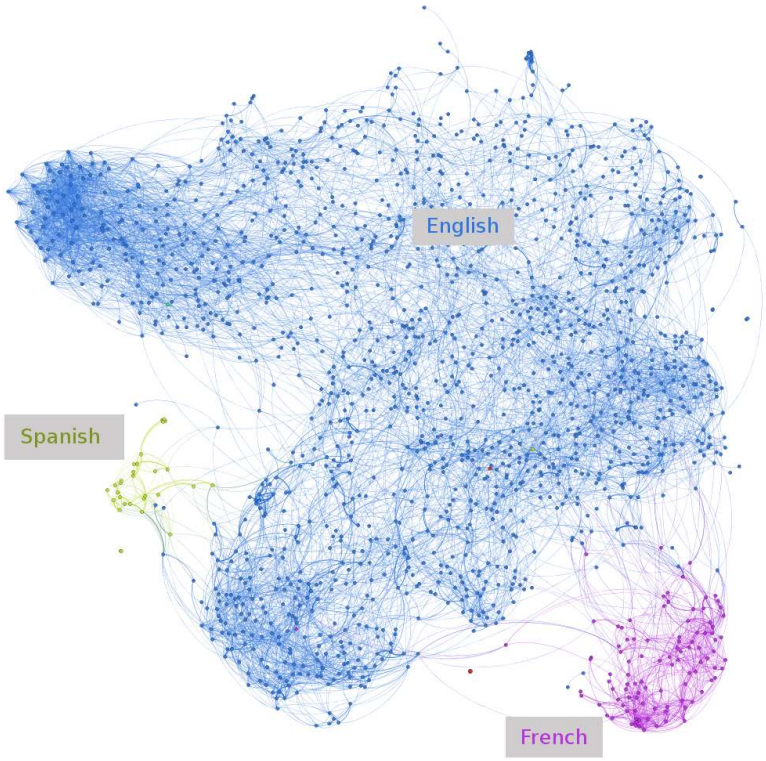
We then load this graph into the graph visualization tool gephi (https://gephi.org/). We layout the graph using the OpenOrd and ForceAtlas2 layout algorithms. (OpenOrd was particularly good at extracting clusters. Beyond that, this was largely a matter of aesthetic taste.)



Graph of Harry Potter Fanfiction (top 1,623 stories)

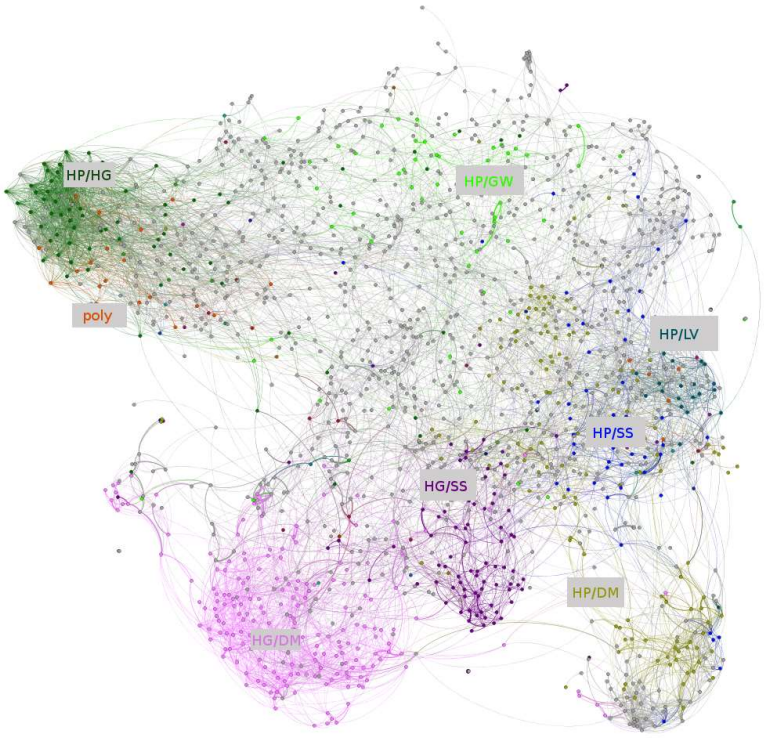
We can see lots of interesting structure in this graph: there are lots of clusters, some more connected than others.

A first hypothesis might be that some of these clusters are caused by language. As it turns out, this is the case:




Graph of Harry Potter Fanfiction, colored by language

Another cause of clusters may be the "ship" (romantic pairing of the story). Many readers have a strong loyalty to a particular ship – for example, they might feel very strongly that Harry and Hermione should be together.

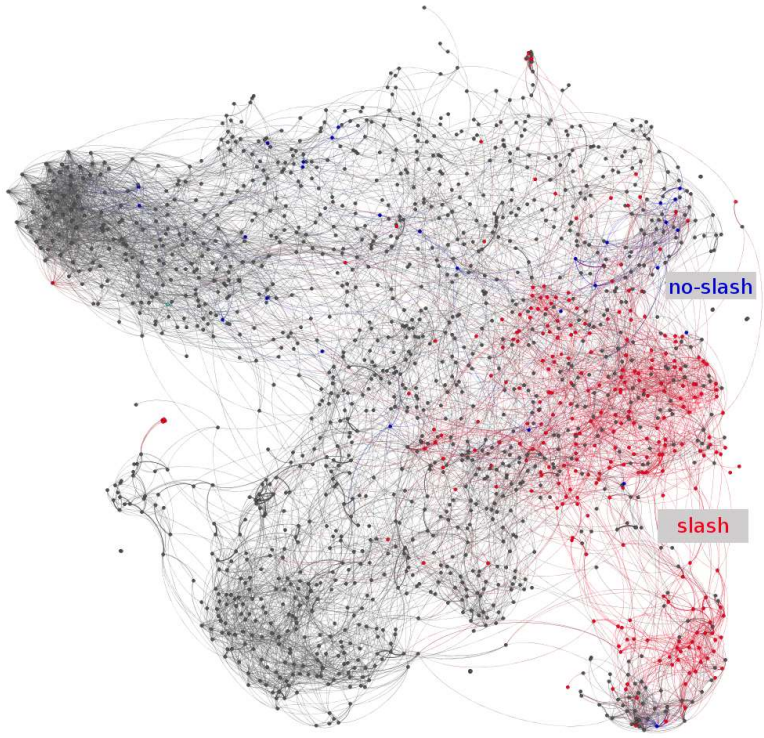



Graph of Harry Potter Fanfiction, colored by ship

(Note: Ships are inferred from tags story summaries. HP = Harry Potter, HG = Hermione Granger, GW = Ginny Weasley, DM = Draco Malfoy, SS = Severus Snape and LV = Lord Voldemort.)

One interesting point is that by far the most diffused ship is HP/GW. It seems likely that this is because it is the ship we see in cannon Harry Potter, and so many stories not focused on romance default to it and unaligned readers are more tolerant of it.

One striking pattern in fanfiction is that a massive fraction of stories are male/male pairings. Such stories are frequently referred to as "slash."




Graph of Harry Potter Fanfiction, colored by slash

Many stories include a slash tag in the summary. Some other stories tag themselves as "no-slash."

One interesting pattern is that stories tagged "no-slash" concentrate around parts of the border of slash stories. One possible reason may be that authors writing stories that might, from a glance at the summary or characters list, look like slash (for example, a story about Snape mentoring Harry, or Draco and Harry as friends) feel the need to explicitly signal that that is not the topic of their story.

The predisposition of the French cluster towards slash stories is interesting, but the cluster is so small I am hesitant to read anything into it.

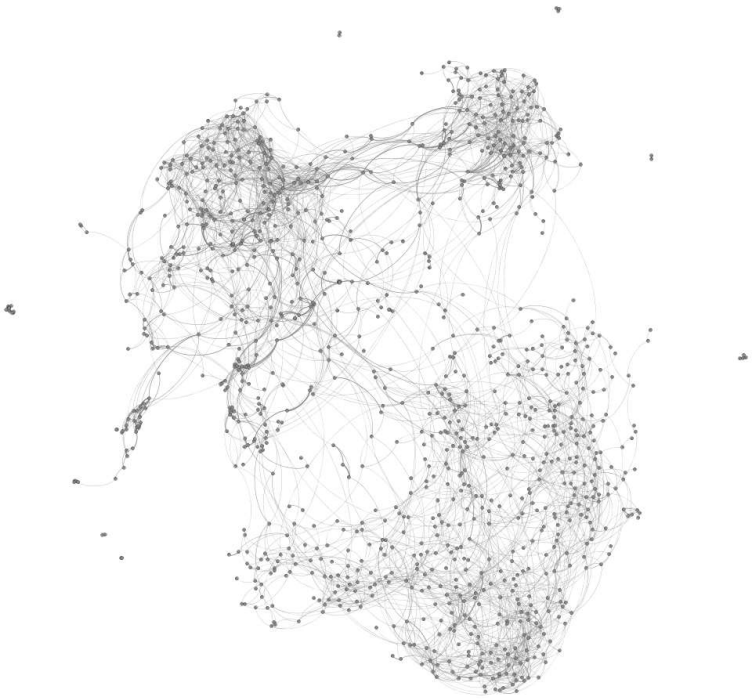
You can also explore an interactive graph of Harry Potter fanfiction (graphs/HP-ship/).

## Large Graph Visualizations for Other Fandoms

Of course, we can apply the exact same tricks to other fandoms.

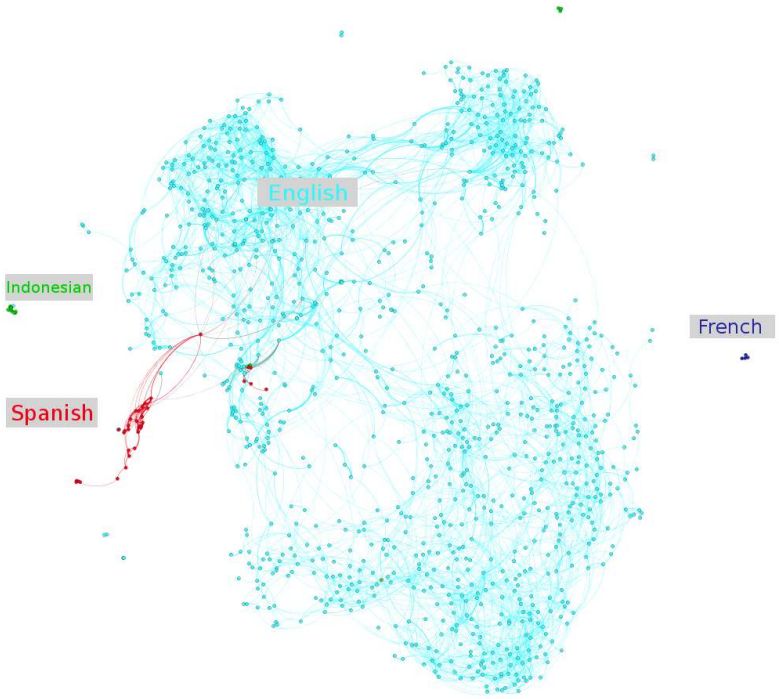
**Naruto**

For example, Naruto is the second biggest fandom. Here's a graph of it:



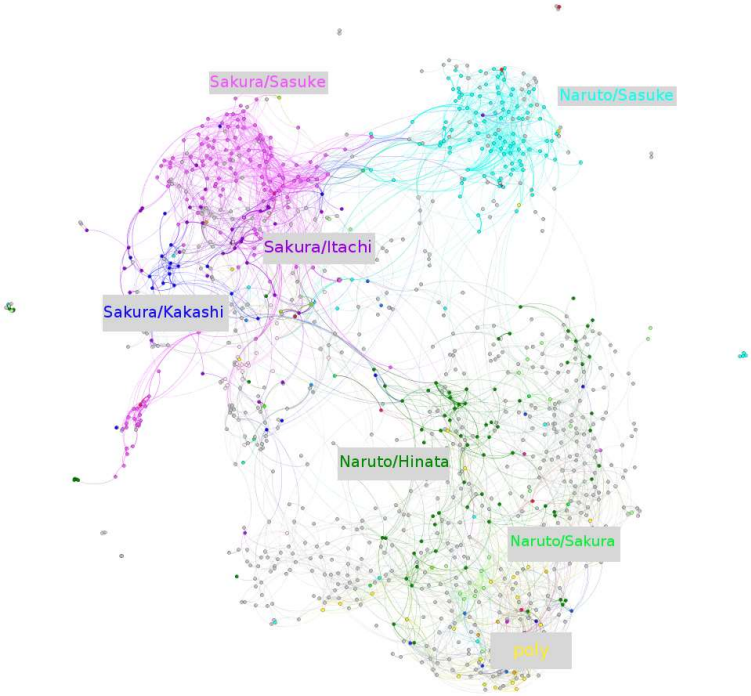
Graph of top Naruto fanfiction (1,123 nodes and 4,000 edges)

We can look at languages again:




Graph of top Naruto fanfiction, colored by language

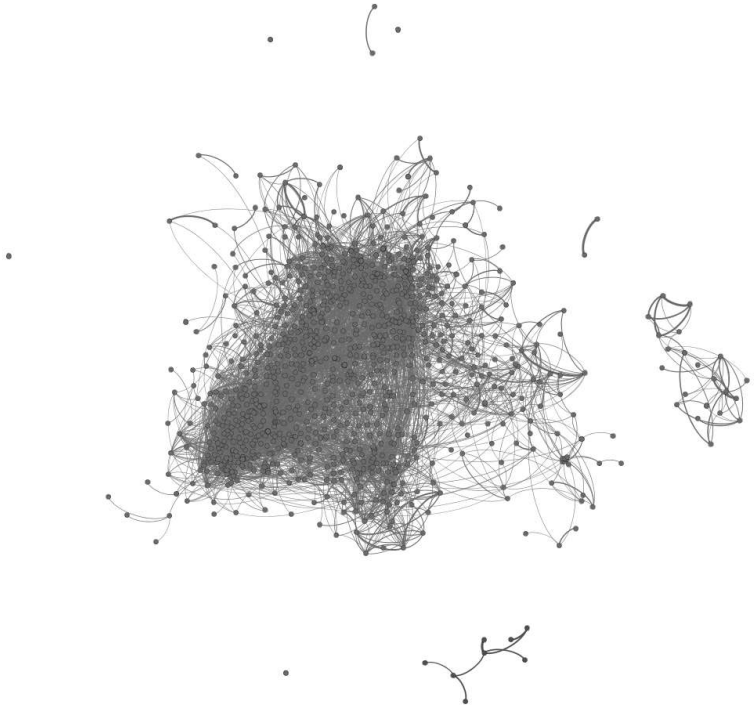
And also for ships:




Graph of top Naruto fanfiction, colored by ship

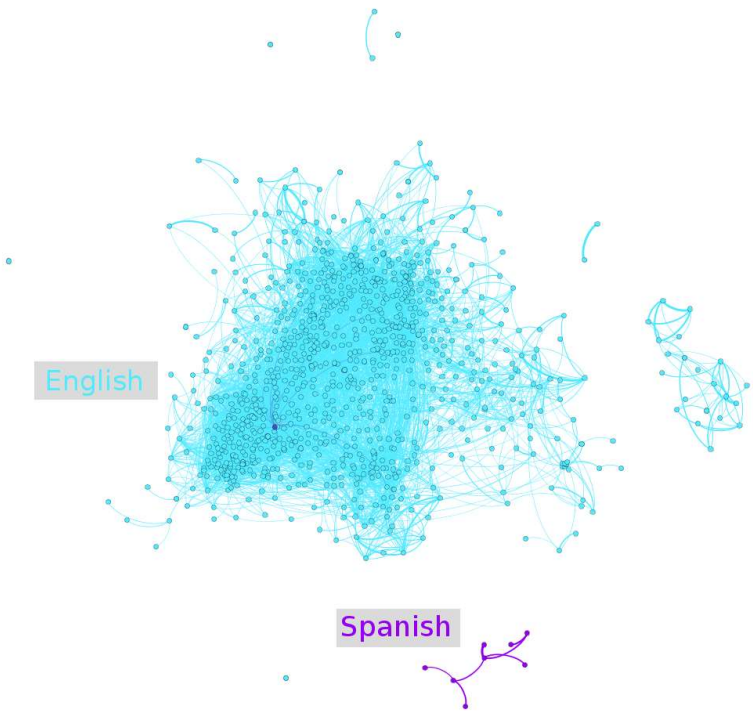
**Twilight**

And again, we can graph the top twilight stories:



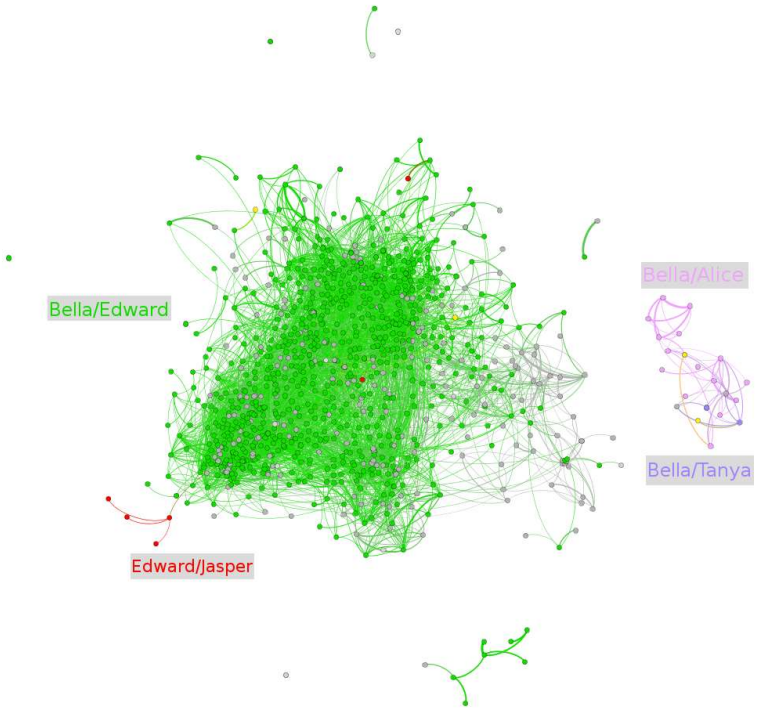
Graph of top Twilight fanfiction (1,031 nodes, 5,00 edges)

We can color it by language:




Graph of top Twilight fanfiction, colored by language

And by ship:




Graph of top Twilight fanfiction, colored by ship

One thing that seems pretty surprising, without inside knowledge of the fandom, is the lack of stories where the pairing involves Jacob. On further inspection, we find that there are stories like that on fanfiction.net, but they aren't amongst the most highly reviewed. Perhaps this pairing prefers other websites? I'd love comments from anyone with insight into this.

You can also explore an interactive graph of Naruto fanfiction (graphs/NAR-ship/) and of Twilight fanfiction (graphs/TWI-ship/).

## PageRank

What are the best fanfics on fanfiction.net? How can we identify them?

A naive approach would be to select the most favorited or reviewed stories. But people's quality of taste varies. A more sophisticated approach is Google's PageRank algorithm which is used to determine which web pages are of high quality.

In a normal vote gives equal weight to every voter. But some voters are better qualified to decide than others. In PageRank, we recalculate the votes again and again, giving each "person's" vote a weight based on how many votes they received in the previous step.

In the case of the Internet, we interpret a website linking to another website as that website voting for the one it links to. Similarly, we can apply it to fanfiction by interpreting story A as "voting" for a story B with a weight of the probability that a user who likes A also likes B.

**Harry Potter top stories by PageRank:**

1. Realizations (http://fanfiction.net/s/1260679) (16.4)

2. Harry Potter and the Nightmares of Futures Past (http://fanfiction.net/s/2636963) (15.7)
3. Make A Wish (http://fanfiction.net/s/2318355) (14.0)
4. Poison Pen (http://fanfiction.net/s/5554780) (11.7)
5. To Shape and Change (http://fanfiction.net/s/6413108) (11.5)
6. **More** (pagerank/hp.html)

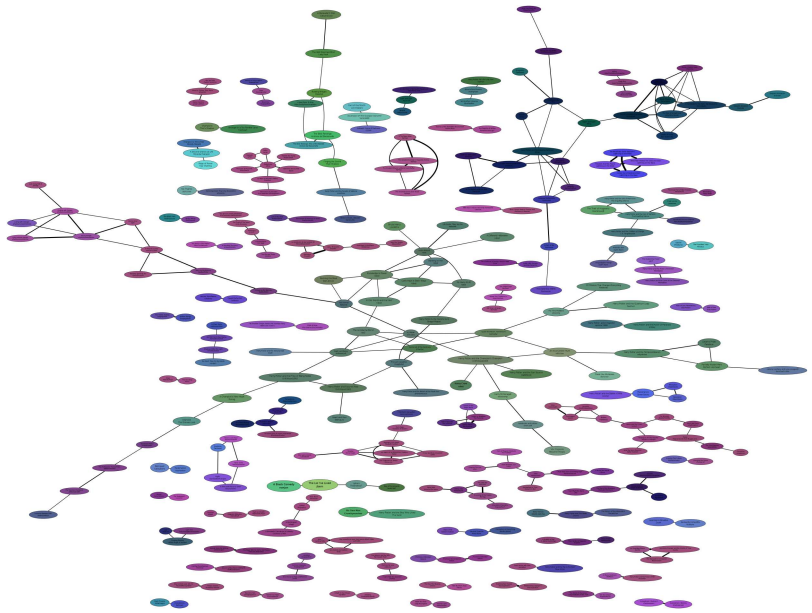
**Naruto top stories by PageRank:**

1. Team 8 (http://fanfiction.net/s/2731239) (11.1)
2. Naruto: Myoushuu no Fuuin (http://fanfiction.net/s/6694302) (6.42)
3. It's For a Good Cause, I Swear! (http://fanfiction.net/s/5409165) (5.57)
4. The Sealed Kunai (http://fanfiction.net/s/6051938) (5.24)
5. Chunin Exam Day (http://fanfiction.net/s/3929411) (5.14)
6. **More** (pagerank/naruto.html)

**Twilight top stories by PageRank:**

1. The Blessing and the Curse (http://fanfiction.net/s/5100876) (18.6)
2. Tropic of Virgo (http://fanfiction.net/s/4901517) (15.0)
3. A Rough Start (http://fanfiction.net/s/5319052) (12.7)
4. Creature of Habit (http://fanfiction.net/s/4769414) (12.6)
5. The Plan (http://fanfiction.net/s/6550419) (10.2)
6. **More** (pagerank/twi.html)

One neat thing we can do is give nodes on our graphs a size based on their PageRank. (We can also color nodes based on the first three components of the singular value decomposition of the adjacency matrix.)



### Story Recommendation

There's something that's just begging to be done, at this point: story recommendations. Given our knowledge of what stories many users like, can we recommend other stories that they're probable to like?

This problem is called collaborative filtering, and is a well-established area. Unfortunately, it isn't something I'm terribly knowledgeable about, so I took a relatively naive approach: sum over the preferences of all users, weighted by how similar their preferences are to the user you are trying to predict.

Specifically, we give each story, $s$, a rank $R_u(s)$, for a user $u$. If the rank is high, we think $u$ is likely to like $s$.

$$R_u(s) = \sum_{v \in F_s \setminus \{u\}} \left( \frac{|S(u) \cap S(v)|}{20 + |S(v)|} \right)^2$$

where $F_s$ is the set of users who favorited $s$ and $S(u)$ is the stories favorited by the user $u$.

For example, we can make recommendations for S'TarKan, the author of the most favorited Harry Potter story on fanfiction.net:

- *Learning to Breathe (http://fanfiction.net/s/2559745) (1.459)
- *Taking Control (http://fanfiction.net/s/2954601) (1.383)
- *Backwards Compatible (http://fanfiction.net/s/1594791) (1.381)
- *Harry Potter and the Nightmares of Futures Past (http://fanfiction.net/s/2636963) (1.377)
- *Harry Potter and Fate's Debt (http://fanfiction.net/s/2479927) (1.218)
- ...

A * denotes that this is already one of the users favorite stories or one of their own stories. We can exclude their favorite stories, and their own stories:

- Make A Wish (http://fanfiction.net/s/2318355) (0.949)
- A Black Comedy (http://fanfiction.net/s/3401052) (0.750)
- Oh God Not Again! (http://fanfiction.net/s/4536005) (0.679)
- Realizations (http://fanfiction.net/s/1260679) (0.642)
- Lord of Caer Azkaban (http://fanfiction.net/s/2107570) (0.635)
- ...

These are all very popular stories. It's not very useful to S'TarKan if we recommend them extremely popular stories that they've almost certainly seen before. As such, it is interesting to penalize the popularity of stories.

Consider $\frac{R_u(s)}{|F_s|^k}$. When $k = 0$, it's our original rank. When $k = 1$, it full normalizes stories against popularity. And in between, it penalizes popularity to varying degrees. If we set k = 0.7, we get these recommendations:

- Insanity (http://fanfiction.net/s/2114122) (0.034)
- Shadow of the Serpent (http://fanfiction.net/s/1995612) (0.032)
- The Bargain (http://fanfiction.net/s/2160456) (0.031)
- Sinners (http://fanfiction.net/s/1975479) (0.029)
- Harry Potter and the Order of the Phoenix (http://fanfiction.net/s/926568) (0.029)
- ...

You can think of these as stories that are *unexpectedly* popular amongst similar users. Similar users like them a lot more than random users like them. (Though, perhaps 0.7 is a bit too extreme.)

Curious about what this algorithm would recommend for you? If you're a popular fanfiction author, you may be in my recommendations for top users for Harry Potter (recs/hp.html), Naruto (recs/nar.html) or Twilight (recs/twi.html).

Since my scripts can't look at your author name while complying with fanfiction.net's terms of service, you will need to know your *author ID*. To get it, go to your fanfiction.net profile page and look at the URL. It will be of the form: `http://fanfiction.net/u/author_ID/...` . Then search for your author ID in the file!

I'm certain one could do much better if they wanted to put a bit more effort into it. :)

## Conclusion

In light of all this, I'd like to reflect on a few things.

**Big Data**: A year ago, I was very dismissive of "big data" as a buzzword. Primarily, it seems to be thrown around by business people who don't really understand much. But one thing I've learned in explorations of data like this one and working in machine learning, is that there is something very powerful about larger amounts of data. There's something very qualitatively different. The fanfiction data I used was actually quite small, only a few hundred users, because of how I limited the amount I downloaded, but I think it still demonstrates the sorts of things that become possible as you have larger amounts of data. (To be honest, a much more compelling example is the progress that's been made in computer vision using ImageNet... But this still influenced my views.)

**Digital Humanities**: Digital humanities also seems to be a bit of a buzzword. But I hope this provides a simple example of the power that can come from applying a little bit of math and computer science to humanities problems.

**Metadata and Privacy**: In this essay, we analyzed stories by looking at whether they were favorited by the same users. There's a natural "dual" to this: analyzing users by looking at whether they favorited the same stories. This would give us a graph of connections between users and allow us to find clusters of users. But what if you use other forms of metadata? For example, we now know that the US government has metadata on who phones who. It seems very likely that many companies and governments have information on where your cellphone is as a function of time. All this can construct a graph of society. I can't really fathom how much one must be able to learn about someone from that. (And how easy it would be to misinterpret.)

**Fanfiction Websites**: I think there's a lot of potential for fanfiction websites to better serve their users based on the techniques outlined here. I'd be really thrilled to see fanfiction.net or Archive Of Our Own adopt some of these ideas. Imagine being able to list a handful of stories in some category you're interested in and discover others? Or get good recommendations? The ideas are all pretty straightforward once you think of them. I'd be very happy to talk to the groups behind different fanfiction websites and provide some help or share example code.

**Deep Learning and NLP**: Recently, there's been some really cool results in applying Deep Learning to Natural Language Processing. One would need a lot more data than I collected, and it would take more effort, but I bet one could do some really interesting things here.

**t-SNE**: t-Distributed Stochastic Neighbor Embedding (http://homepage.tudelft.nl/19j49/t-SNE.html), is an algorithm for visualizing the structure of high-dimensional data. It would be a much simpler approach to understanding the structure of fanfiction than the graph based one I used here, and probably give much better results. If I was starting again, I would use it.

**Resources**: In principle, I'd really like to share my code and make it easy for people to replicate the work I described here. However, I think that would be really rude to fanfiction.net because it could result in lots of people scraping their website, and it seems likely many would remove my rate limiter. An alternative would be to share my extracted metadata, but, again, I think it would be really rude to do that without fanfiction.net's permission, and possibly a violation of their terms of service. So, in the end, I'm not sharing any resources. That said, all of this can be done pretty easily.

*(This post is a fun experiment done primarily for amusement. I would be delighted to hear your comments and thoughts: you can comment inline or at the end. For typos, technical errors, or clarifications you would like to see added, you are encouraged to make a pull request on github (https://github.com/colah/Fanfiction-Graphs-Post). If you enjoyed this post, you might consider subscribing to my RSS feed (../../rss.xml).)*

**Acknowledgments**

Thank you to Eliana Lorch, Taren Stinebrickner-Kauffman, Mary Becica, and Jacob Steinhardt for their comments and encouragement.

## 4 Comments (/posts/2014-07-FFN-Graphs-Vis/#disqus_thread)