Non-Human Social Graphs – a Case Study of MyAnimeList

Wendy Shi, Stanford University
Alfred Xue, Stanford University

In CS224 lecture, we discussed the concept of a small world - the idea that the shortest path between any two individuals is relatively small. This has been attributed to the principle that the likelihood that two individuals know each other correlates significantly with the geographical distance between them. Effectively, this allows each step in the shortest path to travel a percentage of the remaining distance, which causes the shortest path to grow logarithmically with geographical distance, rather than linearly.

We are curious if these principles extend to web networks, which are generally considered to be networks that don't use humans as nodes, but can use their interactions to define edges. In particular, we want to study the degrees of separation between differing Anime, using "myanimelist.net" as our dataset. MyAnimeList (MAL) provides a unique opportunity to take a single set of nodes (anime), and multiple sets of edges that we can analyze. For example, edges can be recommendations from one anime to another, or a shared cast.

Categories and Subject Descriptors: C.2.2 [**Computer-Communication Networks**]: Network Protocols

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Small World, Network Analysis

ACM Reference Format:
Wendy Shi, Alfred Xue, 2016. Diameter of MyAnimeList *ACM Trans. Embedd. Comput. Syst.* 1, 1, Article 1 (October 2016), 6 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

The question

> "If you chose any two people in the world at random, how many acquaintances are needed to create a chain between them?" [Kochen, 1989; Garfield 1979]

describes the small world problem in a succinct manner. Stanley Milgram's famous postcard experiment [citation???] was the first experiment that attempted to measure the degrees of seperation between people in the world. To do so, he selected a broker in Boston to be a target, and had a group of indviduals send leters to acquaintances, seeing if those letters could eventually be routed to the Boston broker. From the results of Milgram's experiment is derived the famous "six degrees of seperation" phrase.

Although there are heavy criticisms regarding the veracity of Milgram's experiments, his work clearly demonstrates two ideas. The first is that the degrees of seperation in a human social graph are far smaller than one would intuitively think, and the other is that the graph is *navigatable*, that is, there exists some algorithm that can traverse from any node s to destination node t only knowing the edges of the current node in $O(\log(N)^B)$ time.

The intuition for both of these properties derives from the idea that the likelihood two people know each other is strongly correlated with their geographical distance. That is, in general, as long as the next hop is geographically closer, then our new node has a shorter path to the destination node than our current one. However, this isn't sufficent to explain the results of Milgram's experiments. For example, if every individual knows every other individual within a one mile radius, but knows no one else, then the shortest path between two nodes eight hundred miles apart requires eight hundred hops – order of magnitues greater than what was observed in Milgram's study. Thus the second requirement is that with high probability, our current node will be connected to another node whose distance to the destination is no more than $\alpha << 1$ times the distance from our current node to the destinatnion node. This allows us to travel to the desitnation much faster – the number of hops now scales logarithmically rather than linerally with the distance to the destination.

Although "six degrees of seperation" is often cited, Milgram's experiment has a few key flaws that should prevent its results from extending to the claim that "all humans are seperated by six degrees." These key flaws are highlighted by Judith Kleinfeld. The first is that both the start and end nodes were recruited from socially active people – basically, in expectation the degrees of seperation between individuals is actually higher, because starting nodes and ending node should be less reachable than the ones used in the experiment. Other criticisms he highlights include that humans have a psychological bias to believe in a short degree of seperation. In general, Kleinfeld criticizes the legitimacy of the study, but does not provide any methods to provide a better study, and his criticisms probably aren't strong enough to reject the concept of a small world, although they are probably strong enough to suggest that six degrees of seperation is insufficent.

Nevertheless, "six degrees of seperation" is probably here to stay, and is supported by more recent analysis on the Facebook social network by Backstrom et. al.

The Facebook network is clearly a social graph, and maintains the properties one would expect from a social graph. Properties such as high clustering, low degrees of seperation, and relatively low connections (i.e. number of edges ¡¡ number of nodes^2) are all observed in this network. Other networks where these properties are observed includes a network of actors, where edges are defined by having collaborated on multiple movies together, or a network of researchers, where an edge is defined to be when two researches have collaborated on a published paper together. The work by Backstorm uses HyperANF, which is similar to what snap.py uses to compute node distances, but uses hyperloglog counters to estimate these distances with high accuracy. HyperANF increases the amount of nodes that can be computed on exponentially with little loss of overall accuracy (individual pairs are less accurate), and allows computation over 700 billion edges.

This study introduces the concept of a spid, or a *index of dispersion* $\frac{\sigma^2}{\mu}$ of the distance distribution. This paper describes the importance of spid as follows,

> "In particular, networks with a spid larger than one should be considered web-like, whereas networks with a spid smaller than one should be considered 'properly social'"

with the logic being that social networks tend to favor short connections (and thus low dispersion), whereas in web/information networks long connections are not uncommon. The results of Backstorm's work clearly indicates that Facebook has a spid well below one. Because this paper used HyperANF, it cannot compute the diameter on the graph – it can only get a lower bound. This is one drawback of using HyperANF, and something to consider when working with large databases. One criticism of this

work is that it does not attempt to draw any meaningful conclusions. It concludes that Facebook is a ”small-world graph” with an average distance of less than 4, but does not attempt to understand any interesting phenomenon observed. We would also liked to have seen analysis on the diameter of Facebook, but that may be infeasible with the computing power at the time.

At this point it becomes clear that there is an interesting distinction between social graphs and web graphs, so we would like to explore the distinction between them. Luckily there has already been work done on this on the twitter graph by Myers et. al. Myers et. al. wanted to determine whether Twitter acted more like a social graph or an information network. The basis of this question is that follows are done through interests rather than relationships, which makes it more like an information network, but initial follows are often based on social interactions. They define the nodes to be twitter acounts and edges to be mutual follows between the accounts.

Myers et. al. also used HyperANF on their dataset of 175 million users and approximately twenty billion edges for mutual follows. (Twitter has more edges per node than Facebook since follows are more common than friendships). In this case, they analyzed multipe graph characteristics of the Twitter graph and compared them to the known characteristics of both information and social graphs. For degre distributions, it is clear that Twitter does not exhibit social graph like characteristics because it has too many individuals who follow and are followed by thousands of people when people are typically unable to maintain more than 150 social relationships at any given time. For both shortest path length and spid, the authors concluded that Twitter exhibited social graph like behaviors, with the shortest path being around 4 (slightly longer than Facebook), and the spid being around 0.11 (slightly wider than Facebook). They also noted that the clustering coefficent of the Twitter graph was relatively high, and thus indicative of a social graph. However, they also noted that the two-hop neighborhoods of Twitter were much much larger than what would be expected in a social graph.

The authors than unhelpfully concluded that Twitter exhibits some characteristics of a social graph and some characteristics of a information graph, but make no attempt to unify the differing characeristics.

It is in this space that we think interesting work can be done.

2. PROJECT PROPOSAL

From the above papers we derive two interesting questions: Are social graphs only demonstrated when nodes are human? Can we use graph characteristics to determine how ”social” different edge types are?

The first comes from the fact that all the papers we have read so far that describe social graphs always use human nodes – people in a friend graph like Facebook, or actors or researchers in a collaboration graph, etc. We are curious if a graph can exhibit social graph features when the nodes are non-human but the edges are derived from humans. One example of such a graph could be research papers that share an edge if they share a mutual author.

The Twitter paper discusses its decision to use mutual follows to generate edges but does some analysis on inbound/outbound only edges. Through this, it becomes clear that (and is rather intuitive) that the sociality of a graph is highly dependent on the type of edges used.

To research both these areas, we propose the following project:

Analyze the graph characteristics of MyAnimeList using different Anime as nodes, and a differing set of edges currently including recommendations and actorship.

The graph characteristics we will analyze are the ones that were used to analyze the twitter dataset, specifically degree distributions, shortest path length, spid, clustering coefficent, and two-hop neighborhoods. We can also add the diameter of the graph,

which was largely ignored by Twitter due to the expensiveness of its calculation. If necessary, we may also analyze subgraphs of MyAnimeList dividing by genre.

2.1. Background on MyAnimeList

MyAnimeList is a website that has aggregate data about tens of thousands of anime. For each anime, it provides metadata information such as genres, ratings, and episode length, as well as voice actors and user recommendations.

There are currently 34240 anime in the database, and around fifteen recommenations per anime, although this varies greatly depending on the popularity of the anime.

There is also user information, where users can input the animes that they have watched and their ratings of them, which can also be used as an edge set, although we have not tried scrapping this yet.

To scrap the Anime data from MyAnimeList, we are modifying an existing python library https://pypi.python.org/pypi/mal-scraper/0.1.0 that no longer works due to MyAnimeList changing their tags.

2.2. Algorithms for Analysis

Luckily, our dataset is not a particularly huge one, so it is unlikely that we will need complex algorithms such as HyperANF to determine our graph characteristics. Degree distribution calculation does not require a particularly complex algorithim, and our graph is small enough that storing a value for every node is not expensive. Both connected component and shortest length pass can be calculated using snap functions. A trivial implementation of two-hop neighborhoods is functionally an $O(VE)$ operation (get set of nodes at one-hop, get set of nodes at two-hop) per node, or a $O(V^2E)$ operation in total. At 30,000 nodes, this is also not expensive. The Floyd-Warshall algorithm can compute the diameter of the graph in $O(V^3)$, which is sufficent for our small graph.

2.3. Evaluation

Like in the Twitter paper, we will compare the graph characteristics of MyAnimeList to a standard social graph to determine which characteristics of a social graph MyAnimeList exhibits. However, we will go one step further and try to explain why only some of the features of a social graph are represented (if that is indeed the case), and attempt to weigh which features are more reflective of a social graph than others. We will then compare the characteristics of MyAnimeList depending on which edge sets we use, and attempt to explain why different edge sets produce different graph characteristics. We can then verify our theories of why edge sets produce specific graph characteristics by perturbing the edge set and then comparing graph charactersitics. For example, if our theory is that MyAnimeList doesn't exhibit a social graph because the average Anime has too many recommendations, we can perturb the graph so that we only include edges between Anime when there have been at least two recommendations between them. We define success to be when we have found a theory of why MyAnimeList exhibits a set of graph characteristics that we can demonstrate the validity of the theory through edge pertubation.

2.4. Submission

At the end of the quarter, we expect to be able to submit our MyAnimeList scrapping code and a paper that both analyzes the graph characteristics of MyAnimeList and provides a theory as to what real-life properties of edges generate what graph characteristics.

3. TYPICAL REFERENCES IN NEW ACM REFERENCE FORMAT

A paginated journal article [Abril and Plant 2007], an enumerated journal article [Cohen et al. 2007], a reference to an entire issue [Cohen 1996], a monograph (whole book) [Kosiur 2001], a monograph/whole book in a series (see 2a in spec. document) [Harel 1979], a divisible-book such as an anthology or compilation [Editor 2007] followed by the same example, however we only output the series if the volume number is given [Editor 2008] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible book [Spector 1990], a chapter in a divisible book in a series [Douglass et al. 1998], a multi-volume work as book [Knuth 1997], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [Andler 1979], a proceedings article with all possible elements [Smith 2010], an example of an enumerated proceedings article [Gundy et al. 2007], an informally published work [Harel 1978], a doctoral dissertation [Clarkson 1985], a master's thesis: [Anisi 2003], an online document / world wide web resource [Thornburg 2001], [Ablamowicz and Fauser 2007], [Poker-Edge.Com 2006], a video game (Case 1) [Obama 2008] and (Case 2) [Novak 2003] and [Lee 2005] and (Case 3) a patent [Scientist 2009], work accepted for publication [Rous 2008], 'YYYYb'-test for prolific author [Saeedi et al. 2010a] and [Saeedi et al. 2010b]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [Kirschmer and Voight 2010]. Boris / Barbara Beeton: multi-volume works as books [Hörmander 1985b] and [Hörmander 1985a].

APPENDIX

In this appendix, we measure the channel switching time of Micaz [CROSSBOW] sensor devices. In our experiments, one mote alternatingly switches between Channels 11 and 12. Every time after the node switches to a channel, it sends out a packet immediately and then changes to a new channel as soon as the transmission is finished. We measure the number of packets the test mote can send in 10 seconds, denoted as N_1. In contrast, we also measure the same value of the test mote without switching channels, denoted as N_2. We calculate the channel-switching time s as

$$s = \frac{10}{N_1} - \frac{10}{N_2}.$$

By repeating the experiments 100 times, we get the average channel-switching time of Micaz motes: 24.3μs.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Maura Turolla of Telecom Italia for providing specifications about the application scenario.

REFERENCES

Rafal Ablamowicz and Bertfried Fauser. 2007. CLIFFORD: a Maple 11 Package for Clifford Algebra Computations, version 11. (2007). Retrieved February 28, 2008 from http://math.tntech.edu/rafal/cliff11/index.html

Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. DOI:http://dx.doi.org/10.1145/1188913.1188915

Sten Andler. 1979. Predicate Path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226–236. DOI:http://dx.doi.org/10.1145/567752.567774

David A. Anisi. 2003. *Optimal Motion Control of a Ground Vehicle*. Master's thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.

Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry)*. Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171.

Jacques Cohen (Ed.). 1996. Special Issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996).

Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalances among conjunctive aggregate queries. *J. ACM* 54, 2, Article 5 (April 2007), 50 pages. DOI:http://dx.doi.org/10.1145/1219092.1219093

Bruce P. Douglass, David Harel, and Mark B. Trakhtenbrot. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. DOI:http://dx.doi.org/10.1007/3-540-65193-4_29

Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:http://dx.doi.org/10.1007/3-540-09237-4

Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. DOI:http://dx.doi.org/10.1007/3-540-09237-4

Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In *Proceedings of the first USENIX workshop on Offensive Technologies (WOOT '07)*. USENIX Association, Berkley, CA, Article 7, 9 pages.

David Harel. 1978. *LOGICS of Programs: AXIOMATICS and DESCRIPTIVE POWER*. MIT Research Lab Technical Report TR-200. Massachusetts Institute of Technology, Cambridge, MA.

David Harel. 1979. *First-Order Dynamic Logic*. Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY. DOI:http://dx.doi.org/10.1007/3-540-09237-4

Lars Hörmander. 1985a. *The analysis of linear partial differential operators. III*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. viii+525 pages. Pseudodifferential operators.

Lars Hörmander. 1985b. *The analysis of linear partial differential operators. IV*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. vii+352 pages. Fourier integral operators.

Markus Kirschmer and John Voight. 2010. Algorithmic Enumeration of Ideal Classes for Quaternion Orders. *SIAM J. Comput.* 39, 5 (Jan. 2010), 1714–1747. DOI:http://dx.doi.org/10.1137/080734467

Donald E. Knuth. 1997. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms (3rd. ed.)*. Addison Wesley Longman Publishing Co., Inc.

David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY.

Newton Lee. 2005. Interview with Bill Kinder: January 13, 2005. Video, *Comput. Entertain.* 3, 1, Article 4 (Jan.-March 2005). DOI:http://dx.doi.org/10.1145/1057270.1057278

Dave Novak. 2003. Solder man. Video. In *ACM SIGGRAPH 2003 Video Review on Animation theater Program: Part I - Vol. 145 (July 27–27, 2003)*. ACM Press, New York, NY, 4. DOI:http://dx.doi.org/99.9999/woot07-S422

Barack Obama. 2008. A more perfect union. Video. (5 March 2008). Retrieved March 21, 2008 from http://video.google.com/videoplay?docid=6528042696351994555

Poker-Edge.Com. 2006. Stats and Analysis. (March 2006). Retrieved June 7, 2006 from http://www.poker-edge.com/stats.php

Bernard Rous. 2008. The Enabling of Digital Libraries. *Digital Libraries* 12, 3, Article 5 (July 2008). To appear.

Mehdi Saeedi, Morteza Saheb Zamani, and Mehdi Sedighi. 2010a. A library-based synthesis methodology for reversible logic. *Microelectron. J.* 41, 4 (April 2010), 185–194.

Mehdi Saeedi, Morteza Saheb Zamani, Mehdi Sedighi, and Zahra Sasanian. 2010b. Synthesis of Reversible Circuit Using Cycle-Based Approach. *J. Emerg. Technol. Comput. Syst.* 6, 4 (Dec. 2010).

Joseph Scientist. 2009. The fountain of youth. (Aug. 2009). Patent No. 12345, Filed July 1st., 2008, Issued Aug. 9th., 2009.

Stan W. Smith. 2010. An experiment in bibliographic mark-up: Parsing metadata for XML export. In *Proceedings of the 3rd. annual workshop on Librarians and Computers (LAC '10)*, Reginald N. Smythe and Alexander Noble (Eds.), Vol. 3. Paparazzi Press, Milan Italy, 422–431. DOI:http://dx.doi.org/99.9999/woot07-S422

Asad Z. Spector. 1990. Achieving application requirements. In *Distributed Systems* (2nd. ed.), Sape Mullender (Ed.). ACM Press, New York, NY, 19–33. DOI:http://dx.doi.org/10.1145/90417.90738

Harry Thornburg. 2001. Introduction to Bayesian Statistics. (March 2001). Retrieved March 2, 2005 from http://ccrma.stanford.edu/~jos/bayes/bayes.html

Received February 2007; revised March 2009; accepted June 2009

Online Appendix to: Non-Human Social Graphs – a Case Study of MyAnimeList

Wendy Shi, Stanford University
Alfred Xue, Stanford University

A. THIS IS AN EXAMPLE OF APPENDIX SECTION HEAD

Channel-switching time is measured as the time length it takes for motes to successfully switch from one channel to another. This parameter impacts the maximum network throughput, because motes cannot receive or send any packet during this period of time, and it also affects the efficiency of toggle snooping in MMSN, where motes need to sense through channels rapidly.

By repeating experiments 100 times, we get the average channel-switching time of Micaz motes: 24.3 μs. We then conduct the same experiments with different Micaz motes, as well as experiments with the transmitter switching from Channel 11 to other channels. In both scenarios, the channel-switching time does not have obvious changes. (In our experiments, all values are in the range of 23.6 μs to 24.9 μs.)

B. APPENDIX SECTION HEAD

The primary consumer of energy in WSNs is idle listening. The key to reduce idle listening is executing low duty-cycle on nodes. Two primary approaches are considered in controlling duty-cycles in the MAC layer.

DOI: http://dx.doi.org/10.1145/0000000.0000000