Thesis draft

Joshua Megnauth

November 4, 2020

1 Introduction: Reddit, gamers, and social networks

Gaming is an omnipresent artistic medium enjoyed by a plurality of Americans. Research from the *Entertainment Software Association* (E.S.A.) has found that about 65% of American adults play video games. About 46% of said gamers are female (Association, 2019). The E.S.A.'s research counters the tired trope of gamers as young, immature boys. Despite the prevalence and diversity of gamers academia is woefully behind on ludology, the study of gaming. Gaming lacks the prestige of other artistic entertainment media such as books, music, and film. An interested party may find texts on the innovations of Allen Ginsberg's poem *Howl* or Sonic Youth's *Daydream Nation* while scarcely finding an academic article on the contributions of the video game *Doom*.

Ludology is as exciting as fields such as A.I. despite lacking the selfsame effervescence. In other words, ludology may be studied from many different angles. Political sociologists would find much to mine from the philosophical schisms in the gaming community. Recent grassroots movements to push for more representation in video games spawned a revanchist and sexist counter movement. Besides the political, sociologists may write ethnographies of the online communities or interactions that occur within gaming. Mark Chen's *Leet Noob: The Life and Death of an Expert Player Group in 'World of Warcraft'* is a study of a group of gamers who tackle dungeons in *World of Warcraft*. Chen writes from the perspective of both a gamer and a social scientist to apply theories such as social and cultural capital, actor-network theory, social construction, et cetera to the video game (Chen, 2011). The list above is clearly not exhaustive. However, I'd like to mention, as an aside, that online communities are transient; ignoring ludology risks permanently missing out on an aspect of culture.

My thesis focusses on the intersection between network science via the social network Reddit and video games. I am particularly interested in the distribution of gamers across subreddits. Reddit and the related terminology will be explained below. I hypothesize that gamers are more likely to post on multiple gaming subreddits (that is, gamers are more likely to have a community that spans across subreddits). Analyzing the network dispersion of gamers is useful for both the social sciences as well as marketing. Network science is concerned with the spread of information by virtue of its focus on ties. Users who post on multiple subs may pass information across subreddits. The graph of gamers may be small world—that is, posters may be relatively well connected to rather than having to "travel" far across Reddit to encounter each other.

My guiding principle is thus to use my statistical and computing skill coupled with my domain knowledge of video games in order to contribute to ludology. Previous studies on gaming lack domain expertise while exhibiting logical fallacies—both of which seem to be caused by the hermetic world of academia. More proper ludology needs to seed the field until the study of video games resembles writing on music or films.

An anecdote engendered my thesis. I've observed; anecdotally, like many others; that nerds of a feather tend to flock together. Gamers tend to share a sample from a common set of interests such as anime, wrestling, technology, certain tastes in music, et cetera. This observation is far from holistic—meaning, gamers do not adopt all of the same traits, and clearly not every person who watches wrestling, for example, is a gamer. However, the basic observation led to my thesis. My research is not seeking to answer that question specifically, but my thesis may provide the groundwork toward studying that larger topic. Gamers may preferentially attach to other gamers or to other ostensibly related subjects such as anime or eSports. Either conclusion—that gamers preferentially attach or not—is useful for ludology.

2 Background: social network analysis

2.1 A digital approach

My goal is to contribute to both ludology as well as studies of Reddit. Gaming by virtue is a heavily technological affair. Gamers are often only lightly insulated from technology. For example, anyone who follows gaming news would likely run into computer terms. Even gamers who do not follow gaming news would run into patches, lag, bugs, or have to know how to connect their consoles to the internet, et cetera. Gaming is unlike television or cable in

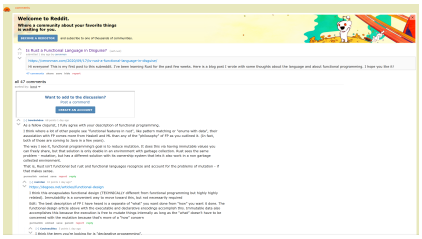
that video games have a strongly "digital native" feel—especially for P.C. (computer) gamers. Gamers discuss video games online on platforms such as Reddit or news sites with forums and comment systems. The culmination of these "techy" qualities does not imply that all gamers are computer scientists but that they are generally reasonably comfortable with technology. All of this is to say that a network analysis focussing on gamers and Reddit is sensible because of the focus on internet media.

2.2 Why video games?

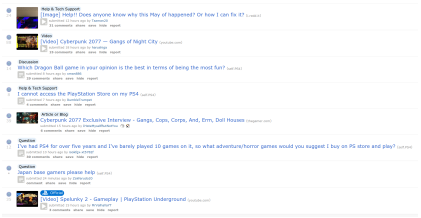
Gamers and ludology, as discussed earlier, are legitimate areas of study that are largely ignored by academia as well as having an image problem regardless of gaming's prevalence. Gaming as a medium is sometimes strangely maligned as coercing individuals toward violence. In 2019, President Donald Trump as well as Kevin McCarthy and Dan Patrick, both Republican politicians, scapegoated video games as reported by an article in the New York *Times*. The article explains that games are often blamed for shootings despite the lack of causal evidence. The American Psychological Association, according to the *Times*, is that video games and violence are not linked. Dr. Chris Ferguson of the A.P.A. humorously mentioned that the data linking bananas to suicide are about the same as those data connecting video games and other media to violence (Draper, 2019).

Most research published on gaming is cursory while also lacking domain knowledge on the subject. A search through databases that compile articles, such as JSTOR, on video games shows that research is decidedly lacking in imagination as they mostly focus on the antiquated violence question. One such article published in *Frontiers of Psychology* exhibits a very flawed methodology in studying gamers. The researchers use an imbalanced and relatively small random sample that skews young and male. Females consisted of a scant 15% of their sample. The writers justified the skewed sample by referencing a two decade old paper despite publishing their article in 2019 and using recent sources otherwise (von der Heiden, Braun, Muller, and Egloff, 2019). Research from the NPD Group corroborates what the E.S.A. found and gamers already know: about half of gamers are women and more females own the 3DS, Wii U, and Nintendo Switch than males (Valentine, 2019). Beyond the poor sampling the research heavily relies on p-values without associated effect sizes. Furthermore, the basis for the study is intrinsically flawed. Gamers are a wide social unit. We would not sample around 3000 music buffs in order to see if "problematic music listening" is associated with "poor psychological functioning."

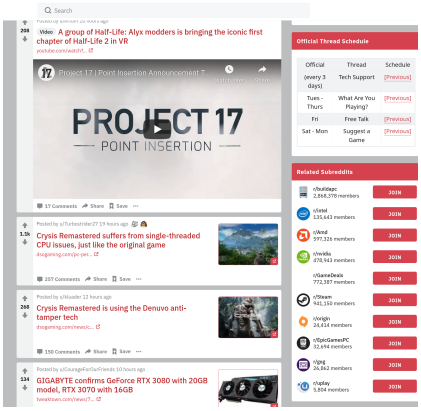
2.3 What is Reddit?



(a) Comments on an /r/rust thread



(b) /r/ps4 threads (old scheme)



(c) /r/pcgaming threads (new scheme)

Figure 1: Examples of Reddit, a forum-like social network

Reddit is a social network reminiscent of the forums and bulletin board systems (B.B.S.) of the past. Reddit is designed for lengthy deliberation, or at least facilitates it, which is like forums and unlike Twitter or Facebook. Subreddits is Reddit parlance for a community which is similar to a forum. Pseudonymous users may post or reply to discussion topics. Like forums, topics may be anything that fits the theme of the subreddit. For example, a subreddit on cats would have topics on cats. Individual subreddits are moderated by screened volunteers from the community; rules are enforced and users may not be able to post if their reputation—known as karma—is too low. Individual topics, known as *submissions*, are also voted up or down. Needless to say, Reddit's structure engenders higher quality content in comparison to Facebook or Twitter—generally speaking at least as Reddit is also home to lengthier discussions on puerile matters. Reddit users, who are pseudonymous, are known by the demonym Redditors.

Reddit reflects gaming's preeminence via *many* the many subreddits on

the topic; individual consoles, series, genres, abstruse memes, as well as technological concepts such as emulation all have incredibly active subs. Reddit is the seventeenth most popular site in the world according to *Alexa*, a subsidiary of Amazon that collects internet data. The social network's position varies from day to day but remains consistently high. Reddit is ahead of several prominent sites such as Netflix, Hulu, Crunchyroll, Instagram, Microsoft's site, Live (Outlook), Twitch, Bing, eBay, AliExpress, and others. Reddit is mainly behind sites such as Chinese social networks such as Weibo as well as Google and YouTube (Alexa, 2020). The largest gaming subreddit on Reddit, */r/gaming*, is the fourth most popular total which places *gaming* above the subreddits for music, sports, relationships, movies, news, gifs, cats, cute pictures, books, programming, et cetera. In fact, the only subreddit more popular than *gaming* that is *not* a default subscription is */r/funny*. The subreddit for *League of Legends*, a single video game, is the sixtieth most popular subreddit (Sizz, 2020).

Redditors may post on as many or as little subreddits as they wish. For example, a user may post on the subreddits for PlayStation 4, P.C. gaming, Linux, and the systems programming language, Rust. Another user may frequent the subs for K-Pop, anime, and the Nintendo Switch. The distribution of these users in terms of where they post is not obvious. In other words, a PlayStation 4 gamer does not necessarily post on the general gaming subreddit nor the subs for other Sony consoles. Redditors who post on multiple subs are part of each community. Thus, the dispersion of gamers, as discussed earlier, is a unique sociological problem. Gamers may be isolated within the specific subreddits they frequent the most or they may be distributed across a wider area.

Reddit is often considered a "news aggregator" like Slashdot, Digg, or StumbleUpon. The latter two are dead while Slashdot, a technology news site, is still alive despite predating Reddit, Digg, and StumbleUpon by over a decade.The researcher personally categorizes Reddit as a forum replacement due to the emphasis on discussions and comments rather than solely existing as curated news (Steinbauer, 2011). However, the reader should understand that Reddit bares a lot of similarities to such services. Reddit's tag line is the "Front Page of the Internet" which implies the service is a portal to content. The content is pulled from a user's own subscriptions to subreddits. Thus, a person may have a front page with the latest political happenings, Linux news, anime discussions, et cetera.

2.4 Very basic network definitions

Network theory focusses on the connections between nodes. Nodes are defined based on context. Nodes in a network of friends would refer to each person. Reddit networks may define nodes as each individual account or poster. Nodes are connected to other nodes via edges. Like node, the term edge is contextual; two friends would share an edge in a friendship network. Edges may be directed or undirected. Consider a video game lending network. The edges may reflect the number of titles loaned. We can define the edge between lender and borrower as directed to account for the relationship where lender is loaning N games out to the borrower.

Graph theorists derive many ideas of relationships as well as formulas and algorithms from this basic concept. We may sum the count of edges a node has to derive degree. The ratio of a node's degree to its potential degree is known as degree centrality (importance). Degree centrality is intuitive because we're calculating a ratio of connections. Thus, more connected nodes would have a higher degree centrality. Nodes traverse paths to reach other nodes. A path exists between two nodes (A, B) if A may cross a sequence of edges to reach B. Additionally, paths only traverse a node or edge once. Betweenness is a metric that measures how often a node is on the shortest path between two nodes. Logically, a node that other nodes often traverse to reach some destination is important in some way (consider airlines or chiefs of staff). A relationship of three nodes creates a triangle or triad. The ratio of triads to potential triads provides a metric for density.

Network theory is obviously more complex than the simple one line, non-mathematical definitions above. However, the basic definitions should suffice for now. Interestingly, some definitions and formulas in the network community seem disputed. Some problems, such as community detection, lack elegant solutions.

2.5 Big data—Volume, Variety, and Velocity

Graph theory does not imply big data, but network science scales well to a high number of observations.Computational social science naturally trends towards big data, such as large graphs or natural language processing. Amir Gandomi and Haider Murtaza argue that the term "big data" is often misunderstood which may impact comprehending a big data project. Big data is traditionally defined by the three Vs: volume, variety, and velocity. Volume is the most straightforward of the three concepts. The "big" in big data suggests a generous amount of information. Volume is simply that—*a lot* of data (Gandomi and Haider, 2015).

Variety and velocity are the more interesting of the three Vs. Big data is often raw and heterogeneous. Unlike the nicely structured flat files data are often distributed as, the majority of existing data are unstructured and multiformat (Gandomi and Haider, 2015). For example, Twitter data would likely arrive in a set of different languages with varying levels of grammar, zeitgeist and context specific references, embedded multimedia, retweets, et cetera. With that said, Twitter data may still be pulled from a clean API which means that Tweets are hardly the most unstructured data a scientist would find. Variety, thus, also refers to the format of data; different image or video formats must be handled, and unstructured data must be appropriately stored as well.

Big data is not only intractably diverse but they are also produced at a high velocity (Gandomi and Haider, 2015). For example, consider all of the data gathered by voice processing systems such as Watson, Siri, Alexa, or Google Assistant. Simply browsing the internet, even with JavaScript and DNS blocking, produces an enormous amount of data. Cellphones collect data more or less constantly, especially with more programs/apps in use. [1]

The three Vs galvanized the use of certain techniques of gathering and analyzing data. Natural Language Processing, graph theory, unstructured databases, clustering, et cetera are not new, but big data have made such techniques both more viable as well as more necessary. NLP, neural networks, as well as computer vision are hotshot fields of data science more generally as well as the computational social sciences due to the prevalence of so much data. After all, what else would we do with said data (Gandomi and Haider, 2015)?

Social network analysis is at least a century old, but big data presents new applications according to Gandomi and Murtaza. Modern S.N.A. is "data-centric" due to the "growing adoption" of social media as well as the many platforms in use. Computational social scientists may analyze either the content or structure of social media data. Network analysis is structural as the *ties* between posters are analyzed. For example, social scientists may use community detection to uncover hidden structures in networks, such as coposters or citation flows for research. Social network analysis may also discover the nodes that are most influential in a network or predict potential links between nodes. Big data's prevalence as well as the massive networks that are common today are driving research into social networks (Gandomi and Haider, 2015).

[1] Data are collected in a manner so megalomaniacal that the European Union instituted the *General Data Protection Regulation* to protect the privacy of E.U. citizens.

2.6 Weak ties in social networks

Mark Granovetter's "Strength of Weak Ties" is a sociological classic that argues that micro level "interpersonal ties" affect macro level phenomena. Granovetter defines a small network consisting of two nodes, A and B, and their friends and acquaintances. Strong ties tend to create other strong ties. For example, A and B's respective close ties would likely meet A or B as well. Nodes in a network transfer information so that our two nodes and their ties gain information from each other. As a small note, Granovetter uses "information" laxly to refer to anything that may be transmitted diffusely, such as knowledge or even disease (Granovetter, 1973).

Information transfer is not limited to strong ties. Granovetter explains that weak ties act like "bridges" that create shortest paths for information to flow. Strong ties act as amplifiers that repeat and retain information. For example, consider A and B's close ties. Rumors, jobs, feats of excellence or failure would all be retained within the strong ties of the network since information is constantly echoed throughout the clique. Weak ties, on the other hand, are "highways" that transmit information rapidly and across distances too. A network of only strong ties is isolated and disjointed. Granovetter uses several examples, including disease; we can consider COVID-19. The coronavirus spread through a series of coincidental ties that eventually made way into areas of stronger ties. In other words, COVID-19 couldn't spread through strong ties alone as that would be too slow. Weak ties diffuse information in general such as jobs, rumors, new ideas, et cetera. Discoveries or new inventions spread easier through networks containing a lot of weak ties due to the heterogeneity of such a network (Granovetter, 1973).

Reddit is strongly attuned to the idea of weak ties. Redditors likely do not post in subreddits consisting of only close ties. Strangers, or at least mild acquaintances, hold topics and discussions on Reddit. Posters may recognize certain Redditors over time, but the connection is still tangential rather than strong. The "velocity" and "volume" of information produced on Reddit facilitates creating weak links as well as information transfer. Some critics contend that the power of weak ties is often exaggerated. Weak ties require some strength behind them to have effect rather than having an effect by default (Kadushin, 2011). Reddit's weak ties have some force behind them due to the focus on discussions. For example, many topics hold in depth conversations on music, films, video games, politics, relationships, computers, et cetera. New games or albums may rapidly traverse the relevant areas of Reddit if popular enough. Such mechanisms operate on Twitter too, of course, but Reddit is specifically geared toward longer communication which in turn holds more power than a couple of characters. Reddit, unlike Twitter,

is not full of fake users and bots that hinder rather than improve connectivity.

3 Structural holes

Ronald Burt extends the weak ties metaphor by arguing that nodes or people near structural holes are more likely to have good ideas. Structural holes are nodes or empty spaces where two different networks meet. Structural holes represent the potential for action. Actors near structural holes may potentially transfer information across networks. Burt argues that the right network conditions must be met with ingenuity. In other words, the bridge or edge must be recognized and then formed. Burt claims his research lacks "novelty" but instead expands ideas that other social scientists have previously laid out in some shape or form. Both Adam Smith as well as John Stuart Mills believed that associating with the "dissimilar" broadens a person's mind. Other social scientists, such as the philosopher Friedrich Hayek, explained how economies integrate over segments (holes) where actors have specialized skills. Burt explains that the general theme of the works of the many disparate writers he covers is that "information, broadly conceived, are more homogeneous within than between groups." Fundamentally, Burt's basic idea is radical for synthesizing ideas previously stated by philosophes like Mills or Hayek into network theory (Burt, 2004).

Consider two networks of corporations where one network consists of food production businesses and the second encompasses data scientists who work in computer vision. A business that resides on the periphery of the two networks or at least has connections to both networks is near a structural hole. The business near the hole may transfer expertise across the networks like Granovetter's bridge edges. Good or novel ideas may arise from the business uncovering areas where the networks may converge—such as using computer vision to improve farming.

The same concept applies to Reddit albeit more abstractly. Structural holes exist near the boundaries of related subreddits. Subreddits do not have *actual* boundaries separating them of course. However, related subs contain information that may be dispersed through weak ties. For example, a user may ask a question about a video game on a general sub that is answered on a more specific sub or vice versa. Users who read both subs are able to direct the questioner appropriately. Readers of multiple subreddits integrate their knowledge into those same subs as well as gain insights from each sub. If the poster writing the inquiry reads multiple subs themselves they will be more likely to encounter an answer. Thus each poster in each sub constitutes a weak tie to each other poster. The potential information available to a

poster in a single sub is thus that of the respondents which in turn spans all of the subs visited by said respondents. In other words, weak ties in subs are potentially projected across all readers. The access to information involves relatively little hops.

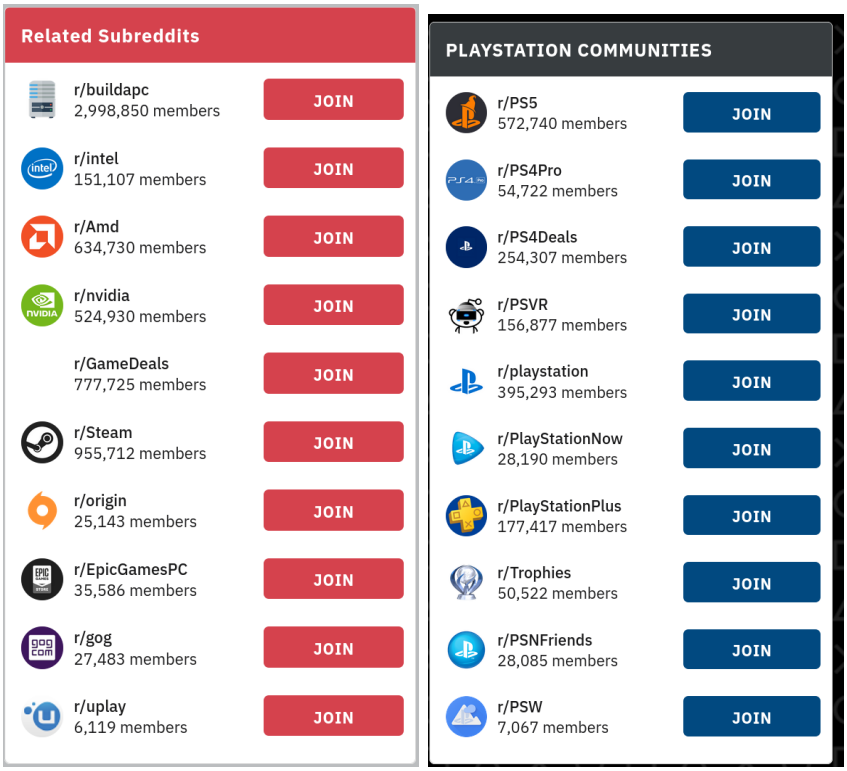
One flaw of the concept above is that we are considering an ideal manifestation of weak ties. Weak ties are important in terms of their bridging capability as per Granovetter. A weak tie may not act as a bridge and thus not transfer information. In our example above, posters who may have answers to questions or tips for the game at hand from other subreddits or experience do not have to share their knowledge. A more nuanced view of weak ties takes into account that posters aren't abstract nodes or machines but must have the willingness to act or the ability to actually arrive at an insight as covered above in the Granovetter section.

3.1 Co-posting networks via Reddit

While studies of Reddit are rather rare, researchers tend to be inclined towards studying *co-posting networks*. Co-posting is essentially a type of propinquity—or residing in the same "space" where space is defined by the topic or researcher. Space in the sense of co-posting refers to residing in the same topic or subreddit. For example, every poster on a specific topic ("I'm struggling to beat x boss. Please help!") would be co-posters. Co-posters may also be defined as all posters on a subreddit such as **/r/KingdomHearts**.

Troy Steinbauer's 2010 paper on Reddit co-posting networks is one of the earliest studies of the social network. Steinbauer gathered submissions and comments from the top 1% of subreddits at the time as well as general data on Reddit. The average subscriber count for subs total at the time of his analysis was about 1077 with a +1 standard deviation of 23514—extremely wide as well as skewed. His research analyzed related subreddits, comment trees, and friend graphs. 2010 is ancient in terms of computational social science, but Steinbauer's methods are still useful for my study (Steinbauer, 2011).

Subreddits may link to other "related" subs as decided by the owners. A related sub is whatever the owners decide. For example, a subreddit on a game series may link the individual subs for each game as well as related series. Steinbauer created a graph where each node is a sub and each edge is a "related" connection between two subs. Subreddits tend to have less than fifteen related links besides a few special subs. The related subs graph is also not very well connected. The largest connected component (LCC) of the graph only contains about 25% of the nodes while the average clustering coefficient is low at 0.039 (Steinbauer, 2011).



(a) Related subs /r/pcgaming (b) Related subs /r/PS4

Steinbauer's comment tree graph is far more connected than than his related subs network. The basis of the network is a set of thresholds for comments. Steinbauer defines a loose graph as having an undirected edge between two posters, A and B, if A or B responded to the other. A tight graph requires both posters having responded to each other. Finally, a strict graph is like a tight graph except with a threshold of four comment pairs between respondents. The final graph is remarkably connected in terms of LCC as the largest component of the loose and tight graphs contain over 98% of the nodes while the strict graph contains about half of the nodes in the LCC (Steinbauer, 2011).

Graham Earley, Nikita Fomichev, Willa Langworthy, and Ruyi Shen studied the Bernie Sanders and Donald Trump subreddits in 2017. *A Statistical Analysis of Network Data From Reddit* is a very mature approach to network science as the researchers use natural language processing (NLP) for some of their node attributes. Earley et alia define nodes as posts to submissions. Particularly, they calculate average sentiment scores per post. The median length of posts is longer on the Sanders sub in comparison to the Trump sub. Posts on the Sanders sub tend to be longer for all three sentiment cat-

egories studied as well—positive, neutral, and negative. Positive posts on the Sanders sub sport an average length of 540.2 compared to 35.36 on the Trump sub. Neutral posts tended to be short on both subreddits (Earley, Fomichev, Langworthy, and Shen, 2017).

Next, Earley et alia calculate network statistics for the two graphs. The Sanders subreddit is much more clustered than the Trump sub at .897 versus .497 for transitivity respectfully. Modularity, or the ability to subset nodes into non-overlapping sets, is also higher for Sanders' sub at .707 in comparison to .505 for Trump. Both of these imply greater connectivity for posters on the Sanders sub as they're responding to each other more across submissions. Average path length is about the same for both graphs despite the difference in clustering (Earley et al., 2017).[2]

So far, we looked at research that considers nodes and edges as Redditors and posts without considering the literal flow of information. Giannis Haralabopoulos, Ioannis Anagnostopoulos, and Sherali Zeadally studied the propagation of information for Reddit, Twitter, and Facebook as well as the media services YouTube and ImgUr. They defined a metric called the Unit of Interest that measured propagation for each service above. The metric varied per service as each defines "success" differently. For example, Facebook's metric was defined as likes per post. The gathered data were filtered to preclude posts that failed to double the UoI four times. Gaming ended up as one of the most prevalent topics, especially on Reddit, after the data were filtered (Haralabopoulos, Anagnostopoulos, and Zeadally, 2015).

3.2 Assortativity and co-posting networks

Assortativity is the tendency for nodes to be attracted to other nodes with similar characteristics. Both "nodes" and "characteristics" are general here—assortativity (also known as homophily) is not limited to social science. Networks with high assortativity tend to be more robust to removing nodes and edges as theorized in a seminal paper by M.E.J. Newman. Social networks tend to have higher assortativity than other types of networks. Humans nat-

[2]Reddit recently banned **/r/the_donald** due to users spearheading attacks on other communities. The community's notoriety spans beyond coordinated trolling. The sub bred conspiracies and racist memes. The posters eventually fled to an echo chamber and "alternative" Reddit that respects their first amendment rights. Note that Reddit isn't a governmental organization that can violate the freedom of speech (Tiffany, 2020). I assume that the relatively disconnected nature of the sub shown in the paper may be due to the sub lacking the cohesiveness of an actual community. The Donald is more of an ideology, like *juche sasang*, which in turn means the sub was closer to the puerile subs noted earlier. Users may have visited and posted memes and rants rather than engaged in substantive conversation.

urally flock together for certain qualities such as, perhaps, politics. Computational social scientists have studied graphs of online social networks, such as Twitter, to determine if robustness holds for assortativity. The paper *Co-posting Author Assortativity in Reddit* is one of the first to see if Newman's theory holds for Reddit (Cauteruccio, Corradini, Terracina, Ursino, and Virgili, 2020).

Research by Cauteruccio, Corradini, Terracina, Ursino, and Virgili tested Newman's hypothesis for Reddit. The researchers gathered 150,795,895 observations that they used to create a co-posting network. Density is low at 0.006 which is sensible considering the wide range of data scraped. Cauteruccio et alia first calculated assortativity based on degree centrality. Posters were binned into intervals 312,500 nodes wide. The researchers calculated assortativity for the extreme intervals as well as the intermediate intervals. After comparing their results with a null model created with permutation the researchers found that posters in an interval are more likely to have more connections within their interval than outside (Cauteruccio et al., 2020).

3.3 Community mapping and small world networks

Computational social scientists may map out ephemeral communities using networks. Communities and roles may be *emic* or *etic*. *Emic* refers to nominal position assigned by a group such as president or manager. *Etic* refers to the actual structure as discovered by some process such as network science, anthropology, political science, et cetera (Kadushin, 2011). For example, an analysis of the United States government would show the relative power of the nonpartisan independent agencies that execute the law. Analyzing Reddit may reveal communities that span borders (subreddits).

The digital humanities are known for having a fuzzy definition and community due to the diverse fields of study. Martin Grandjean applied social network analysis to Twitter in order to determine the shape of the community. Grandjean gathered the users from Tweets mentioning specific hashtags, keywords in biographies, Twitter's search functionality, and volunteers who had heard of his research (Grandjean, 2016).

The final graph is a low density network that nevertheless exhibits small world qualities. Most users are individually weakly connected. In other words, users may have access to the entire network while not being deeply rooted in it. Language created clear clusters in the network despite interlanguage connections. Languages present in Grandjean's data include English, French, German, Dutch, and Spanish. Grandjean found that users with high betweenness centrality reside at the intersection between these communities (Grandjean, 2016).

Self-similarity and small world are often considered two antagonistic properties. Small world graphs may not be fractal—the mathematical property of repeating as length increases. However, complex networks, such as social networks, may be self-similar depending on the subset. Gallos et alia studied a subset of the IMDB movie network that consisted only of adult film actors. The subset allowed them to use a relatively small and focussed data set that was also small world. The researchers set a threshold of co-starring connections in order to get rid of connections that obscure fractality. Some connections are relatively less important than others. Therefore, precluding those weak connections in order to reveal possible fractality is an acceptable trade. The researchers implemented a box covering algorithm that showed network fractality for the reduced graph when only stronger ties were considered (Gallos, Potiguar, Andrade, and Makse, 2013).

3.4 Information spread in online social networks

Information spread is something of a hot topic in terms of studying online social networks. Virality, the rise of social movements, the spread of propaganda or conspiracy theories such as QAnon, memes, information dispersion, et cetera are all phenomena that may be studied in terms of social networks. Sen Pei et alia studied determining superspreaders in networks. Superspreaders amplify messages which in turn mean they're of high importance for everything from marketing to social movements (Pei, Muchnik, Andrade, Zheng, and Makse, 2014).

The researchers found that algorithms like Google's PageRank perform poorer than expected. Pei et alia argue that prior papers are flawed due to the way they simulate network data to test algorithms like PageRank. These papers simulate data based on epidemiology via random walk models. PageRank, which is similar to eigenvector centrality, tends to perform well with the simulated data. However, different algorithms outperform PageRank depending on the problem. K-core performs better than PageRank for real world epidemiology data but poorly on rumor data, for example. Furthermore, simulated models fail to handle properties such as assortative mixing which occurs naturally in real world networks (Pei et al., 2014).

K-core outperformed other algorithms based on the findings by Pei et alia. Metrics such as K-core or eigenvector centrality are global metrics that require complete networks to be properly calculated. Studying online social networks such as Reddit, Twitter, or Facebook is thus extremely difficult or perhaps impossible unless a clear subset is demarcated. Thus, they collected the entirety of the LiveJournal network in order to both circumvent the errors caused by simulated data as well as the difficulty of studying influence

in other online social networks (Pei et al., 2014).

Gathering the entire LiveJournal (LJ) network allowed the researchers to track information diffusion globally. In other words, the researchers were able to follow every link made by a user to an LJ blog other than their own. Starting with a user, i, the researchers followed the chain of users who linked to posts by i as well as users who linked to the links et cetera. The method accounts for every user rather than conferring importance on whichever users the algorithm picked as a starting point. The result is the influence of every user whose content was shared. K-core outperforms PageRank which works via a random walk. Random walks may work for ranking website links but performs suboptimally for social networks (Pei et al., 2014).

4 Methodology

```
2020-09-19T09:59:19.002Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T09:59:32.722Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Scraped 100 nodes from https://api
.pushshift.io/reddit/comment/search?sort_type=created_utc&subreddit=PS3&size=1000&sort=desc&before=1600382930.
2020-09-19T09:59:32.724Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T09:59:46.253Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Scraped 100 nodes from https://api
.pushshift.io/reddit/comment/search?sort_type=created_utc&subreddit=PS2&size=1000&sort=desc&before=1600286005.
2020-09-19T09:59:46.254Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T09:59:58.502Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Scraped 100 nodes from https://api
.pushshift.io/reddit/comment/search?sort_type=created_utc&subreddit=PSX&size=1000&sort=desc&before=1600189647.
2020-09-19T09:59:58.504Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T10:00:13.376Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Scraped 100 nodes from https://api
.pushshift.io/reddit/comment/search?sort_type=created_utc&subreddit=emulation&size=1000&sort=desc&before=1600285827.
2020-09-19T10:00:13.377Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T10:00:24.729Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Scraped 100 nodes from https://api
.pushshift.io/reddit/comment/search?sort_type=created_utc&subreddit=otomegames&size=1000&sort=desc&before=1600380344.
2020-09-19T10:00:24.731Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T10:00:35.936Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Scraped 100 nodes from https://api
.pushshift.io/reddit/comment/search?sort_type=created_utc&subreddit=demonssouls&size=1000&sort=desc&before=1600428089.
2020-09-19T10:00:35.937Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T10:00:47.549Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Scraped 100 nodes from https://api
.pushshift.io/reddit/comment/search?sort_type=created_utc&subreddit=metalgearsolid&size=1000&sort=desc&before=160038852
4.
2020-09-19T10:00:47.555Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T10:01:11.330Z ERROR thesis_gamer_scraper::scraperclient::scraperclient > Reqwest: error decoding response b
ody: EOF while parsing a value at line 1 column 0 @ https://api.pushshift.io/reddit/comment/search?sort_type=created_ut
c&subreddit=gaming&size=1000&sort=desc&before=1600445951
2020-09-19T10:01:11.330Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Sleeping: 10 seconds
2020-09-19T10:01:24.788Z INFO  thesis_gamer_scraper::scraperclient::scraperclient > Scraped 100 nodes from https://api
.pushshift.io/reddit/comment/search?sort_type=created_utc&subreddit=LeagueOfLegends&size=1000&sort=desc&before=16004452
10.
```

Figure 3: Scraper hard at work

My research project requires relatively recent data as well as observations relevant to video games. Researchers in the computational social sciences seem to trend toward gathering their own data. The internet is a dynamic space of constant data production. Structured data may be limited or flawed for many topics, including ludology. Scraping data allows computational social scientists to gather raw, unpasteurized data rather than relying on old, static sources. A survey, for example, is saliently flawed for gamers because conducting one properly would likely require a large sample across a wide span of ages as well as capture gamers who play on different platforms.

Scraping data seems both more reliable as well as more direct in a case like that as well as for my project.

While large Reddit data sets exist via the *Stanford Network Analysis Project* as well as *Kaggle*, scraping data from Reddit is not too difficult due to the open Reddit API (Reddit, 2020). Bindings for Python exist via *PRAW* as well. However, the Reddit API imposes several restrictions and requires an account to use properly as well. The restrictions are reasonable due to Reddit's high traffic: rapacious scrapers may hammer the site and associated Content Delivery Networks. Caching services, such as the open source Pushshift, aims to alleviate the stress caused by scrapers as well as provide a more convenient API (Baumgartner, n.d.). I opted for Pushshift to gather my data due to its flexibility.

My scraper gathers N Node-Edge pairs from a provided list of subreddits. I wrote the scraper in the systems programming language, Rust, and the source code is available on GitHub[3]. My program's high level logic is as follows:

1. Collect 100 nodes with associated metadata from a subreddit. Each observation contains the Redditor name, epoch timestamps, topic, and subreddit name.

2. Paginate by setting the *before* API parameter to the earliest timestamp from the recent scrape of 100 nodes.

3. Add each gathered node to a hash set which precludes duplicate observations.

4. Repeat steps one through three for the set of subreddits in a round robin manner.

5. Repeat steps one through four until N is reached.

6. Remove deleted users and automoderator replies.

7. Hash all of the Redditor and topic names with SHA256 at an attempt at privacy[4].

8. Deserialize everything to a CSV.

[3]https://github.com/joshuamegnauth54/thesis_gamer_scraper

[4]Very little of the data sets found online actually hashed the user names as the information is all public. However, hashing at least adds a minimum layer of privacy.

My data set[5] contains about 127,000 observations gathered via the logic above. The first and most salient limitation here is that I only gathered 127K observations for a select set of subreddits. I originally intended to perform a breadth first search by scraping outwards from the nodes. So in that case each node in the set of unique nodes would be individually queried for the list of subreddits they posted to at least N times to add noise to the data. Ironically, Pushshift lacks an easy access point (as far as I know) for gathering that data. My program contains the requisite logic to scrape that data from Reddit without using their API[6], but Reddit bans users who are caught scraping data outside of their public API.

My analysis uses the open source Python[7] library NetworkX (Hagberg, Schultz, and Swart, 2008). Python is an interpreted, high level general purpose programming language. In short, Python is powerful and fast to write. NetworkX is a native Python library that implements an impressive collection of graph algorithms including random graph generators. Alternatives include igraph, a C library with Python and R bindings; and petgraph, a Rust crate. NetworkX sports an easy and pragmatic API but is also far slower than libraries written in speedy systems languages. Besides NetworkX, GraphViz and Pandas, both open source, were used for auxiliary tasks such as visualizations as well as wrangling my raw data. The source code for the analysis is located in the same repository as the thesis itself.

References

Alexa. (2020). Alexa - Top Sites. Retrieved from https://alexa.com/topsites

Association, E. S. (2019). 2019 Essential Facts about the Computer and Video Game Industry. Retrieved from https://theesa.com/esa-research/2019-essential-facts-about-the-computer-and-video-game-industry/

Baumgartner, J. (n.d.). Pushshift API. Retrieved from https://pushshift.io

Burt, R. S. (2004). Structural Holes and Good Ideas. *The American Journal of Sociology*, (2).

Cauteruccio, F., Corradini, E., Terracina, G., Ursino, D., & Virgili, L. (2020). Co-posting Author Assortativity in Reddit. *CEUR-WS*, (14).

Chen, M. (2011). *Leet Noob: The Life and Death of an Expert Player Group in 'World of Warcraft'*. Peter Lang Inc., International Academic Publishers.

[5] Available here: https://github.com/joshuamegnauth54/GamerDistributionThesis2020/data

[6] See: https://github.com/joshuamegnauth54/thesis_gamer_scraper/blob/master/src/scraperclient/scrapercli The scrape_individ_users function implements the logic.

[7] https://python.org

Draper, K. (2019, August 5). Video Games Aren't Why Shootings Happen. Politicians Still Blame Them. *New York Times*.

Earley, G., Fomichev, N., Langworthy, W., & Shen, R. (2017). A Statistical Analysis of Network Data From Reddit.

Gallos, L. K., Potiguar, F. Q., Andrade, J. S., & Makse, H. A. (2013). IMDB Network Revisited: Unveiling Fractal and Modular Properties from a Typical Small-World Network. *PLOS One*.

Gandomi, A. & Haider, M. (2015). Beyond the Hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Grandjean, M. (2016). A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, (3).

Granovetter, M. S. (1973). The Strength of Weak Ties. *The American Journal of Sociology*, *78*, 1360–1380.

Hagberg, A. A., Schultz, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*.

Haralabopoulos, G., Anagnostopoulos, I., & Zeadally, S. (2015). Lifespan and propagation of information in online social networks: A case study based on reddit. *Journal of Network and Computer Applications*, (56).

Kadushin, C. (2011). *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press.

Pei, S., Muchnik, L., Andrade, J. S., Zheng, Z., & Makse, H. A. (2014). Searching for superspreaders of information in real world social networks. *CUNY Academic Works*.

Reddit. (2020). Reddit API. Retrieved from https://reddit.com/dev/api

Sizz, M. (2020). Retrieved from https://redditlist.com

Steinbauer, T. (2011). Information and Social Analysis of Reddit.

Tiffany, K. (2020). Reddit is Done Pretending The Donald is Fine. *The Atlantic*.

Valentine, R. (2019). EEDAR: Nintendo Switch Attracting More Women, Wider Age Ranges Over Time. Retrieved from https://www.gamesindustry.biz/articles/2019-02-11-eedar-nintendo-switch-attracting-more-women-wider-age-ranges-over-time

von der Heiden, J., Braun, B., Muller, K., & Egloff, B. (2019). The Association Between Video Gaming and Psychological Functioning. *Frontiers in Psychology*.