Modeling the Relationship of Game Streamers and their Viewers

Nicolas Homble
Johns Hopkins University
Baltimore, MD, USA
nhomble1@jhu.edu

Ian McCulloh
Johns Hopkins University
Baltimore, MD, USA
Ian.mcculloh@jhuapl.edu

John Piorkowski
Johns Hopkins University
Baltimore, MD, USA
jpiorko2@jhu.edu

***Abstract*--- Use of generated media from web sites like YouTube and Twitch are a growing source of much of the media people consume today. This has become an important medium for companies to target their advertisements and sponsorships. This can take the form of content providers explicitly mentioning their sponsors in their content, or adding company branding and merchandise into the content. These are familiar tactics in any celebrity-like testimonial. What is interesting on sites like YouTube and Twitch is that there are explicit metrics around the number of impressions these medias produce. With the social network backing the viewers and content authors, there are social forces that can promote linkage between consumers and authors which could lead to more targeted advertising efforts. With this context, we looked at the live broadcasting network Twitch.tv network for analysis. To quantify these relationships, we used exponential random graph models which has been shown to be an appropriate model for social networks. We can show that invariant properties of other social networks appear in the Twitch gaming network and homophily is a significant force in network linkage.**

I. INTRODUCTION

Sharing and consuming content from other online users of your choice is a basic hallmark of a social network and is a major focus for advertisers. Many studies have looked into the predicting consumption of products and services based on social media data successfully [1]. In parallel to the rise of social media content and the interest to study it, "esports" or competitive gaming has steadily been rising in popularity and revenue. While isolated to particular cultures and games (take StarCraft in Korea), the sport recently boomed within the past decade. While the term began for Arcade tournaments in the 80s and multiplayer games started to form in the 90s, it was in the early 2000s where leagues like Major League Gaming were formed which led to the community growing to 205 million [2].

Twitch in the past decade has grown into a global, user-driven media content platform where users are able to broadcast their own custom channels live for anyone to watch online. This is very similar to one of the earliest social networks from the dot com bubble with Josh Harris's pseudo.com where users could webcast for other users. Instead of going bankrupt though, Twitch has also grown to become one of the largest drivers of internet traffic by attracting millions of users onto the platform. While beginning as primarily a video game broadcast platform, Twitch.tv does not define creative boundaries on content apart from legal and common decency measures. Users in the Twitch community can both create live content in their own channel (these users are often called streamers) and other users can join the channel anonymously, can follow with no payment obligations, or can subscribe and pay periodic payments to support their favorite streamers.

This background is important because it sets the stage for the importance in the analysis of the Twitch social network. There is interest and money in the games that are played. Game developers who can generate a competitive league like "League of Legends" or a community like "Minecraft" to create loyalty within gamers and pave the way for more lucrative opportunities through merchandising and events. Popular games then make for interesting content on sites like YouTube and Twitch. It then stands to reason that interesting statistics about games can be derived from online consumption. Additionally, for the streaming community, this is becoming a viable occupation for many with top streamers earning millions of USD annually. It is interesting to understand what factors promote linkage between streamers and their viewers [3].

Among other social network platforms, Twitch is especially interesting for analysis because of its open public application programming interface (API). Additionally, there are stronger directed ties between streamers and users through the "follow" function in Twitch. These details among other characteristics, make Twitch a unique social network for study that is interesting and accessible for data scientists. To be brief, there are a couple important components of the Twitch system. First, streamers have the same user accounts as any other viewer and they are able to follow and be followed like everyone else. This easily contrasts with YouTube where a content provider is more often seen as a repository for video resources; in this case, streamers are both a generator and active participant in the social network. In terms of content, users can both live stream and they can store static videos for offline viewing (typically highlight reels). Second, Twitch focuses more on live broadcasts rather than static offline content. This leads to metadata around content to be more ambiguous as long running streams are closer to variety shows rather than topic specific video. Finally, the majority of Twitch content is focused around gaming. While there is art, music, and just chatting related channels, the bulk of content on Twitch revolves around gaming. This sets the focus for our research. We are interested in understanding this gaming network as we hope relationships will help shed light into targeted advertising and product initiatives.

Given the context of Twitch's social network and the use cases, we arrive at the general question we'd like to answer: Can we model the behavior of Twitch viewers? This is broken down into three questions we will attempt to tackle in this report:

Q1: Do streamers drive new content in the community or does the community force streamers to migrate to new games?

Q2: Does the Twitch network present previously demonstrated social network invariant traits?

Q3: Can we identify nodal attributes that promote linkage between users?

These questions will be explored with data sampled from the Twitch v5 API. Looking at both network central tendencies and statistical graph models, we will aim to find interesting relationships.

II. BACKGROUND

Related Work

There has been extensive work in both understanding the popularity of user generated content and surveying massive social networks to understand the relationship between users in those networks. Both give a proper overview of the sampling bias that can occur when attempting to pull a manageable dataset from these massive networks. Every second, for example, YouTube is growing its catalogue in the order of days [2]. Data collection then becomes an important topic for every paper in order to find bias. The authors in [2] were able to use YouTube's APIs to find recent and videos using hardcoded keywords provided by the researchers. In [5], network traffic to public sites were sniffed on campus to simulate the real time data flow. While other authors have also looked into non-mainstream video sites like DailyMotion, little work seems to have been done towards a broadcast site like Twitch.tv.

Along with extensive work done analyzing similar social networks, common network invariants have been repeatedly noted in social networks. While network relationships can be difficult to speak broadly against, especially because of domain specific entities, a frequently cited network invariant is that the in-degree of nodes demonstrates a power loss function. Additionally, network homophily has been shown to lead nodes in a network to visit content that their nearby links go to [5].

Outside of academia, there was an interesting article created by the Twitch data science division in 2015 that did a quick survey of time users spent in different communities in the network. Among several implications, we can derive several social factors in their visualization [4]. We expect homophily to play a large role in this network since we see nodes under certain Twitch communities well connected. The question for to answer in this paper is to see if similar traits of homophily can be derived across streamers on the platform and if relationships can be found through other means than private data like user viewing times.

Network Theory

There has already been work into the properties of various types of complex graphs. Social networks as well have been studied and typically exhibit certain properties which we plan to relate in section IV. Here we discuss some of these common properties.

Social networks commonly conform to a *power-law* network, where the probability of a node having a degree k is a function of $k^{-\gamma}$ for large k and a constant tuning parameter $\gamma > 1$. Previous work has demonstrated that social networks are of this class of networks [5].

Exponential Random Graph Models

Exponential random graph models give us a means of quantifying and understanding structural dependence in our social networks. Considering the probability of any two nodes connected in a network, we find that the probability of such a connection cannot be viewed independently from the rest of the network. The goal then is to a fit a probability function that describes linkage between nodes. Given an observed network as a start which provides attributes and a number of edges, an exponential random graph routine should look parameters to fit our model that maximizes likelihood of our model from our initially observed network.

The general form for exponential random graph models takes the form of [6]:

$$\Pr(\mathrm{Y} = \mathrm{y}) = \frac{1}{\kappa} e^{\sum_A \eta_A g_A(y)}$$

Where A represents the set of all graph configurations, which leads η to be the parameters corresponding to configuration of A and g(y) then ties to the provided statistic to the model for that configuration. Finally, κ is a normalizing constant so that we have a probability distribution (between 0 and 1). This is quite a general form and leaves open the statistic used which in our case would be attribute matches across nodes in our networks. Additionally, we have left open how we solve for such a model.

For a realistic model, Markov dependencies are introduced to model the fact that a relationship between two nodes is affected by the relationships those nodes have with others. In a social network, the intuition is there. Then due to complexity of tracking the inter dependencies, we arrive at a pseudo-likelihood estimation of the model [6]:

$$log\left[\frac{\Pr(Y_{ij} = 1|{y_{ij}}^C)}{\Pr(Y_{ij} = 0|{y_{ij}}^C)}\right] = \sum \eta_A d_A(y)_{A(Y_{ij})}$$

All observations in graph excluding a specific observation between I and j are captured in the complement, ${y_{ij}}^C$. In our work, we will be using the *statnet* [7] package.

III. DATASET

Sampling the Twitch Network

In the case of most interesting, massive social networks, it can be infeasible to analyze the entire networks due to size; therefore, some method of sampling must be employed. Unlike in [5], sniffing network traffic was infeasible for a multitude of reasons. Therefore, the data collected for this paper was done through API requests to Twitch's v5 service. This meant we were limited to asking for information about a certain user, which includes who they follow and who follows them. From the user object, we are also able to obtain a list of static video content with associated metadata such as game name and creation date. One limitation, however, is the ability to gain random content from the service. In [1], the authors could rely on recently added material. Unfortunately, there is no equivalent API from Twitch. The provided search API allows us to choose some hardcoded query string which we substituted by hardcoding central nodes to the Twitch network.

This left us with some sampled graph traversal. Prior papers consistently mention that interrupted breadth first searches over a social network would skew the data. In the case of Twitch though, we think there are reasonable assumptions we can make on the network which makes preserving depth not as important. We hypothesize Twitch is driven by users following streamers (users), streamers may have an interesting network but users should be close if not already terminal nodes on the graph. We also assume that a consumer that produces no content would have no other followers from the community. Then we should expect reliable depth and breadth coverage if we recursively search from hand chosen popular root nodes and scale the breadth search by the number of in and out edges to that node. By scaling the search, we aim to reduce skewing of nearby nodes, and we are confident the width of our sample will reflect the actual user base.

Twitch Metadata

To briefly summarize the data we can call from Twitch's API,:

- For a user, you can query for their followers and who they are following
- For a user, you can query for videos uploaded by them
- For a video, you can ask for metadata like the game name to correlate with game metadata
- For a game, you can ask Twitch for the current number of viewers and "popularity" score

Then to answer questions such as, what game does this streamer typically broadcast, we look at their 30 day video history and pull the most frequent game name.

Constructed Networks for Analysis

We now have sufficient data collections for various views of analysis. From our general sampling we are able to compute high level statistics of the Twitch network and we can verify common social network traits. For the network relationships, we needed to minimize our dataset even further. To address Q3, we further constrained our network sampling to users with degrees on the polar ends of the spectrum. We separated users we deemed as streamers based on some number of followers being less than some parameter *N*. With these two groups, two networks were derived for analysis. To look at streamer-to-streamer relationships, a network was created only linking user with *totalFollowers* > *N*.

For the second network, we still linked streamers together however an edge existed between two streamers if there existed a viewer that followed both streamers. This was an interesting view since it gave a glimpse into whether users tended to follow games or streamers.

Then along with the adjacency matrix to form the network, the attributes of the nodes were a combination of game titles found in the user's broadcast list and the fact that a user followed a certain organizational user. An example would be player who followed the official channels for certain esport games like StarCraft, League of Legends, and DOTA. Additionally, we found interesting results in the streamer-to-streamer network by checking for specific links between users and organizational users in the network like game companies and competitive leagues.

IV. ANALYSIS

High Level Statistics

From our sample, we can look for some basic trends in the network. Our sample pulled 290 thousand user profiles as a base before any filtering was done. We see the activity by popular streamers tends to reflect on the community's popularity of the game.

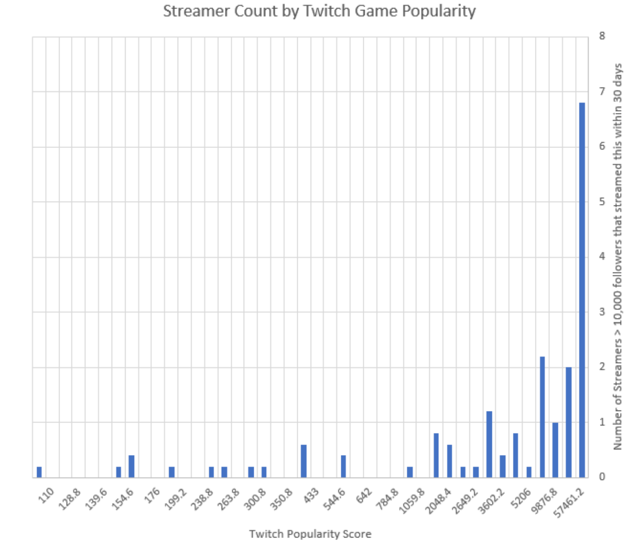

Figure 1: number of sampled streamers with more than 10,000 followers who play games of rising popularity defined by Twitch

This by itself is not too surprising in that the number of content providers should meet the demand; however, it is unclear if the content created by streamers is a result of user demand or users are interested in what the streamers produce.

We can also look at the creation date of popular users versus the general population.

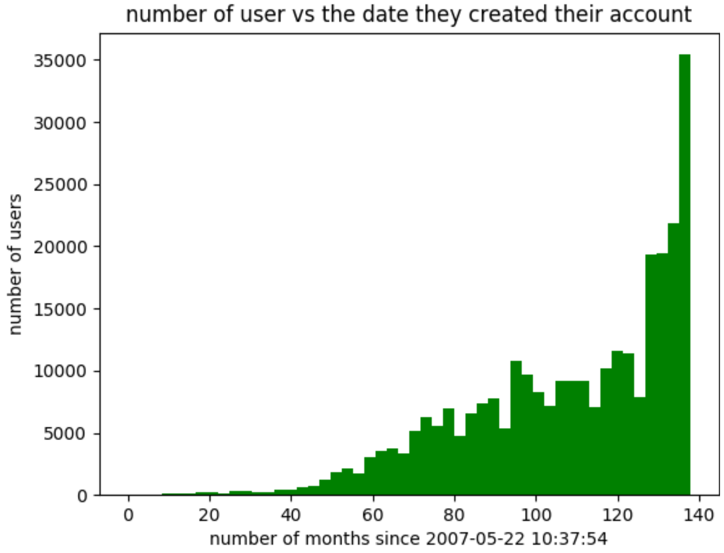

Figure 2: number of created sampled accounts over time

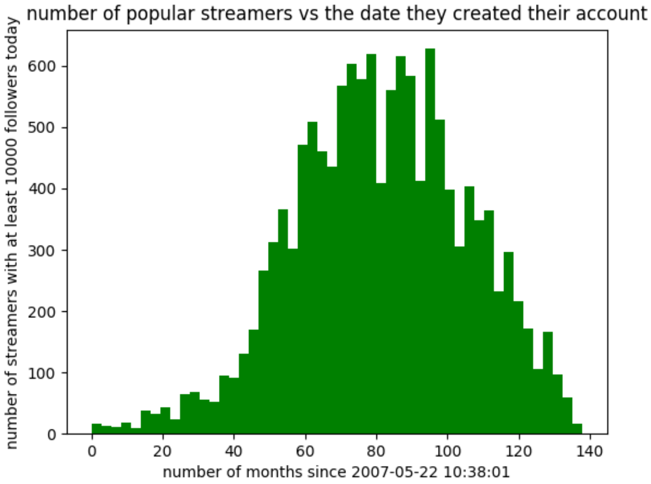

Figure 3: number of created sample users with at least 10,000 followers as of December 2018 over time

We can clearly see an explosion in the growth of Twitch's user base in the past couple months but popular streamers seemed to follow a more normal distribution centered around 80 months after May 2007 which evaluates to January 2014.

This would imply that streamers are not able to immediately capitalize on the wave of new games otherwise we would expect a spike in the popular streamer accounts by game releases such as *Fortnite* and *Players Unknown: Battlegrounds*. We can then partially answer Q1 in that streamer popularity cannot be entirely reactionary to new game titles; however, without more temporal data it is hard to understand how existing popular streamers migrate to new content.

Twitch Social Network Properties

To explore Q2, we look beyond simple trends in our general data pull, we can also examine common social network traits in our data. As was mentioned previously, we expect the degree distribution of our network to follow a power-law. Given Twitch's data model, the in-degree is simply the number of total followers a user has. Because of the disparity between users with no followers and a few followers, the following plot has the frequency on a log scale.

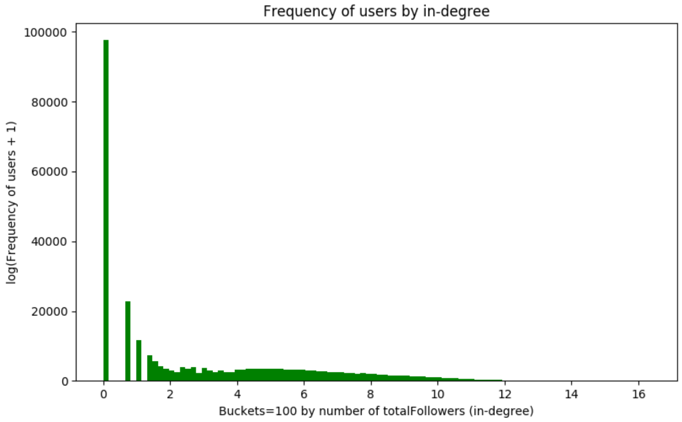

Figure 4: The number of users bucketed by increasing in-degree from 0 where the y-axis is on a log scale

The exponential drop off is clear. Similarly, we can measure the frequency of users by out-degree by looking at the total number of users a Twitch user follows. Of the 200 thousand users, there were about a thousand users who followed more than 1800 users so they were removed from the following plot as outliers. We then see an extremely nice drop off as defined by a power-law:

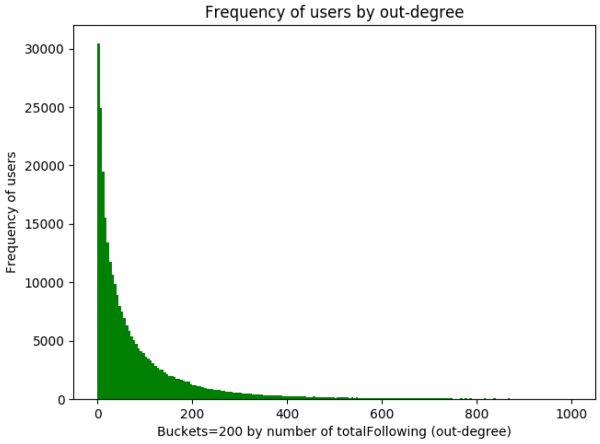

Figure 5: The number of users by increasing out-degree from 0

Note that this y-axis was not a log scale like the in-degree plot. Based on the properties described by [5], we see the drop off in node degree like other studies social networks exhibit. Not only does this give some confidence in the sample taken for this study, but it also reassures that social media analysis techniques to be used in section F will have merit.

Figure 4 and 5 also give a sense of information distribution in the network. From work in [5], we finally look at the proportion of out to in degree.

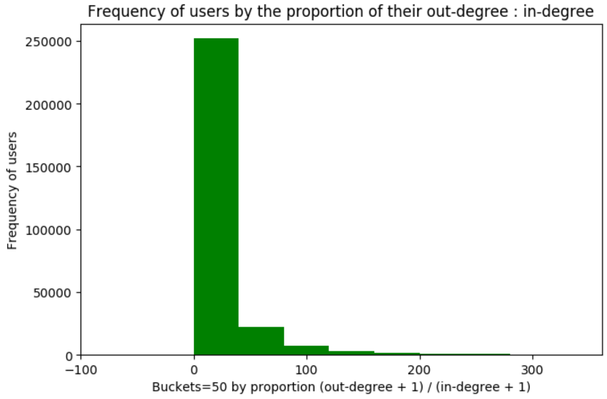

Figure 6: frequency of users across the proportion of out to in degree

The trend towards a higher in-degree than out-degree that also drops offs very quickly seems to suggest that the network of users without a popular following is dwarfed by the popular following of a smaller group of users. This seems to the 1-way flow of information we would expect on a video sharing site like Twitch and confirms that the network in Twitch is possibly driven by elements of popularity and prestige. Unlike a more extensive network where we want to relay information, the Twitch network is driven by users following their favorite gamers.

User Relationships

To answer Q3, we will use the exponential random graph technique described earlier to identify node attributes that increase the probability of an edge between users. First, we consider attributes that lead streamers to follow each other. Our intuition leads us to think that similar gameplay will lead users to follow each other since it is not uncommon for games to feature each other and even partner together. Instead, we found that users grouped better together by following the organizational units behind the games they were playing.

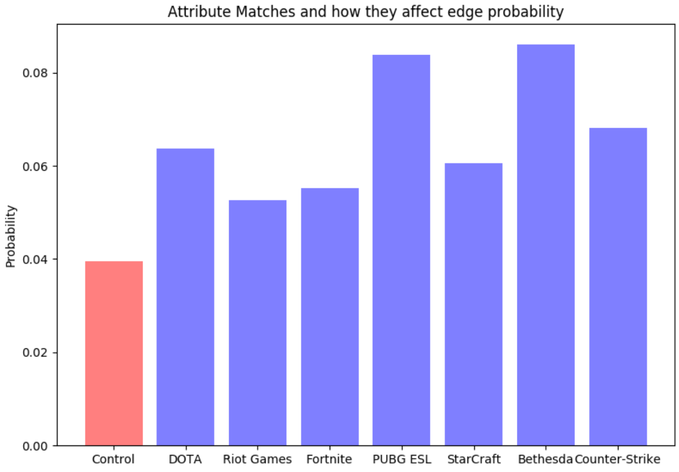

Figure 7: Edge probability between streamers across different node attributes

Against the control probability where no node attribute constraints were added, we can see probabilities even double for users who follow the *Players Unknown: Battlegrounds* league channel. Of course, one might suspect that the relationships are driven by games with an esport following, but we see even a game company like *Bethesda* can be an indicator for streamers following other streamers.

To further investigate user interaction, we consider factors that would lead viewers to follow another streamer. Given the limited about of metadata for a viewer with no generated content, we examine if viewers follow other streamers who create similar content.

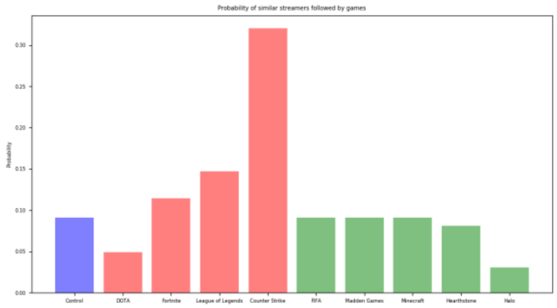

Figure 8: Edge probability between two mutually followed streamers by game they both stream

Interestingly, we find that despite some unifying factors amongst content providers, the attributes that lead a user to follow to users is highly variable. Given the control probability, we see that in competitive games like *League of Legends* or *Counter Strike: Global Offensive* there is sometimes a high probability that a follower will follow multiple users streaming the same content. For non-competitive games such as *Minecraft* or *Halo* which has a smaller esport community, we see no real change or even a decrease in correlation.

V. Discussion and Conclusion

Key Findings

In this paper, our goal was to better understand the Twitch network by looking at various statistics, network properties, and evaluating different node attributes to understand network relationships. We have first shown that the Twitch network shows elements of a social network which opens the door for more study of this platform like other content sharing networks. The in and out degree of the network demonstrate the common features of a social network with a power loss function. This sets the stage for continued study of this network like other content sharing social networks. We obtain a view of the general structure of the network to be a prestige driven network where actors typically follow less people than their total number of followers. This feeds into the impact companies can see from targeting advertisements and sponsorships on select users. Additionally, we see that the platform has greatly grown in the past couple months which adds to the importance of understanding this network along with its peers.

Finally, we can confirm insights given by Star in [4] and we can demonstrate elements of homophily among streamers through a random graph model by looking at affiliation to game organizations. This is an important confirmation for streamers and companies alike because this makes the case for companies to support the growth of their game community through an official channel on these sites. This supports the idea that to create a community, it helps to have some kind of official structure or organization for users to rally against. This also does not assume we are only considering esports.

Limitations

Twitch is a flexible platform that does not require streamers to fit a particular mold. Within this paper, we tried to associate users with a certain set of content, but the reality is much more complex. Users are left to their creative devices. Anecdotally, the top streamers are not just great gamers. They are entertainers. This means that live broadcasts are not necessarily a showcase of a particular but can even be devoted to just chatting to viewers. Because of the flexibility, this means that metadata around user streams are less than plentiful.

Additionally, we were left data available for public consumption which hides many network relationships that would give more meaningful relationships. For example, following a user requires no effort by the viewer and may not reflect the consumer’s current view of the streamer. In contrast to subscriptions on the Twitch network, users pay small periodic feeds to be subscribed to a streamer. Given the monetary requirement, one can expect more up-to-date state of interest between users.

VI. Future Work

The data collected in this paper was a snapshot in time on a small sample in comparison to the size of the community. Future studies should learn more about predicting future behavior by looking at user decisions and networks as they evolve over time.

Additionally, content generated on Twitch is frequently exported to other mediums. While Twitch prides itself in providing live gaming content, highlight reels are made viral on other social network sites like Reddit and YouTube. It would be an interesting study to tie actions and users from the Twitch network to their presence and content in these other sites as well. Content on YouTube, for example, might be richer with video metadata and similar videos that the platform natively provides. Similarly, relating written reactions on Reddit to content created or seen on Twitch would add a new dimension of attributes for network analysis. In our current study, the only action a listener could do in our network was to follow a streamer. If we consider users of Reddit consuming the content, there is richer information on context on that subreddit thread that could be exposed via natural language processing techniques.

Overall, there is exciting work to be done in this area with a growing user base in parallel to a growing field. The esport community and general video game industry continues to grow in capitol with an expectation to reach 138 billion USD at the end of 2018 [8]. The company that can tap into these streamer networks to showcase their game or game merchandise stands to join the growth of this new business domain.

REFERENCES

[1] S. Asur and B. Huberman, "Predicting the Future With Social Media," arxiv.org, 2010.

[2] A. H. Olsen, "The Evolution of ESports," Coventry Univ, Coventry, 2015.

[3] T. Kim, "Tyler 'Ninja' Blevins explains how he makes more than $500,000 a month playing video game ‘Fortnite’," 19 March 2018. [Online]. Available: https://www.cnbc.com/2018/03/19/tyler-ninja-blevins-explains-how-he-makes-more-than-500000-a-month-playing-video-game-fortnite.html. [Accessed 1 December 2018].

[4] E. Star, "Visual Mapping of Twitch and Our Communities, ’Cause Science!," 4 Feb 2015. [Online]. Available: https://blog.twitch.tv/visual-mapping-of-twitch-and-our-communities-cause-science-2f5ad212c3da. [Accessed 1 Dec 2018].

[5] A. Mislove, P. Druschel, M. Marcon, B. Bhattacharjee and K. Gummadi, "Measurement and Analysis of Online Social Networks," in *SIGCOMM,* San Diego, 2007.

[6] D. Lusher, J. Koskinen and G. Robins, "Exponential Random Graph Models for Social Networks," Cambridge University Press, Cambridge, 2013.

[7] C. f. S. i. D. a. Ecology, "statnet," [Online]. Available: https://statnet.org/. [Accessed 1 December 2018].

[8] K. Ell, "Video game industry is booming with continued revenue," 18 July 2018. [Online]. Available: https://www.cnbc.com/2018/07/18/video-game-industry-is-booming-with-continued-revenue.html. [Accessed 1 December 2018].

[9] "twitch.tv Traffic Statistics," Alexa, [Online]. Available: https://www.alexa.com/siteinfo/twitch.tv. [Accessed 1 Dec 2018].

[10] G. Robbins, P. Pattison, Y. Kalish and D. Lusher, "An introduction to exponential random graph models for social networks," *Social Networks,* no. 29, pp. 173-191, 2006.

[11] "Exponential Random Graph Models (ERGMs) using statnet," 2014. [Online]. Available: http://statnet.csde.washington.edu/workshops/EUSN/current/ergm/ergm_tutorial.html. [Accessed 1 December 2018].