**Key Business Problem**

**Twitch:**

* Twitch is a live streaming platform that primarily focuses on video gaming, including live streaming of gameplay, esports tournaments, and gaming-related talk shows.
* Twitch allows viewers to watch live or on-demand content, interact with streamers through chat, and participate in the community.
* Twitch's business model is multifaceted, combining advertising, subscriptions, donations, and sales of digital content.

**Context:** In the rapidly growing live streaming industry, retaining streamers (especially those who are actively contributing to the platform's content diversity and quality) is crucial for sustaining platform growth, viewer engagement, and overall revenue.

**Problem Statement:** Twitch faces the challenge of streamer churn, where content creators stop actively streaming or leave the platform altogether. This could be due to various factors, including low viewer engagement, insufficient monetization, or the allure of competing platforms offering better incentives. Churn not only impacts Twitch's content quality and diversity but also affects viewer retention and the platform's attractiveness to advertisers and sponsors.

Notably, no prior research has employed network analysis to tackle the problem of churn prevention, with traditional methods focusing primarily on the analysis of individual attributes. Yet, the influence of social networks is significant, manifesting in scenarios where individuals cease their engagement with the platform following their peers' departure, highlighting the need for a network-centric approach to understanding and mitigating churn.

|  |  |
| --- | --- |
| A graph showing a line of growth  Description automatically generated with medium confidence | <https://www.statista.com/statistics/746173/monthly-active-streamers-on-twitch/>  Number of active streamers on Twitch worldwide from January 2018 to January 2024 (in millions)   * Statista provides comprehensive data and insights into Twitch's active streamer count, among other platform statistics, such as the reported 8.36 million active streamers in January 2024, down from a peak of 9.89 million in January 2021 |

**Proposed Network Solution Strategy**

**Overview:** Leveraging the extensive social network and feature data from Twitch, we propose a machine learning-based solution to predict potential churn among streamers. Our strategy involves identifying key predictors of churn and engaging at-risk streamers with targeted interventions to retain them on the platform.

**Dataset**:

The dataset consists of a social network of Twitch users collected via the public API in Spring 2018, featuring nodes representing Twitch users and edges indicating mutual follower relationships.

* 168,114 nodes
* 6,797,557 edges

User categories:

* **~~Regular Users~~**~~: These are individuals who use Twitch to watch content and follow other users. They may not necessarily create content themselves.~~
* **~~Broadcasters~~**~~: These are users who actively stream content on Twitch. They can be regular streamers or those with Affiliate or Partner status. The dataset includes attributes like broadcaster language which are specific to users who create and stream content.~~
* **~~Affiliates and Partners~~**~~: This subset of broadcasters have met certain criteria set by Twitch and have access to additional monetization features. The dataset includes the 'Affiliate Status' attribute which helps identify these users.~~

1. Large Twitch Edges (large\_twitch\_edges.csv)

|  |  |
| --- | --- |
| Column | Definition |
| source | The ID of the user who is following another user. |
| target | The ID of the user being followed by the source user. |

2. Large Twitch Features (large\_twitch\_features.csv)

|  |  |  |
| --- | --- | --- |
| Column | Definition | Values |
| ID | Unique identifier of the Twitch user. |  |
| Dead | Indicates whether an account is inactive (dead). | 1 or 0 |
| Language | The broadcaster's primary language. | EN, ES, … |
| Affiliate | Whether the user has affiliate status on Twitch. | 1 or 0 |
| Mature | Indicates if the user streams explicit content. | 1 or 0 |
| Creation Date | The date the account was created. |  |
| Last Update | Last date the account information was updated. |  |
| Views | Total view count of the user's channel. |  |
| Account Lifetime | The length of time the account has been active. (days)  Account Lifetime = Days between Creation Date and Last Update. |  |

**EDA**:

* Draw the graph >> whether a cluster is at high risk of churning
* How to split the dataset into train and test

**Model**:

Input:

* **Profile Features:** Language, affiliate, mature, account­ lifetime in the large\_twitch\_features.csv.
* **Social Connectivity Features**: from the large\_twitch\_edges.csv

|  |  |  |
| --- | --- | --- |
| **Measurement** | **Definition** | **Implication for Churn** |
| Degree Centrality | Number of followers | This is a strong indicator of a streamer's popularity and the ability to attract and retain an audience. It directly correlates with potential revenue streams (e.g., subscriptions, advertisements, donations) and engagement levels, making it a valuable input for predicting churn among broadcasters. |
| Degree Centrality 2 | Number of mutual followers a user has on Twitch | This is a critical metric for understanding the strength of a user's social network and engagement within the Twitch community. High mutual followership indicates strong reciprocal relationships, which can be a significant factor in a user's decision to stay active on the platform. It is highly suitable for models that aim to predict churn by analyzing social integration and community support aspects. |
| Betweenness Centrality | How often a user acts as a bridge in the shortest path between two other users | High values indicate a user's significant role in connecting different parts of the network, suggesting active engagement that might reduce churn. Low values might signal isolation or peripheral network position, possibly increasing churn risk. |
| Closeness Centrality | The average length of the shortest path from a user to all other users in the network | High closeness means a user can quickly interact with others, which could signify active engagement and lower churn risk. Low closeness might indicate greater network isolation, potentially increasing churn risk. |
| PageRank Centrality | A variation of Eigenvector Centrality, considering the network structure and the importance of nodes | High PageRank indicates a user is considered important within the network structure, likely reducing churn risk. Low PageRank might suggest a less central role, potentially increasing churn risk. |
| Clustering coefficient | The extent to which a user’s followers are also followers of each other | High clustering coefficient suggests a tight-knit community, possibly enhancing user engagement and reducing churn risk. |

* Why not number of followings: not directly link to churn. A high value might indicate the user is actively exploring OR less focused and transient use of the platform.

**End goal**

* **Importance analysis**: explain which is the major factor and advice how to retain customer (business suggestions)
* **Causality analysis** (potential analysis) - Highest churn score >> whether they will affect people around them or they are affected by others around them

Output:

* **Dead**: binary 1 or 0

Proposed Network Model:

* Logistic Regression, random forest …

Table

|  |
| --- |
| Dear team,  Good work on the pitch yesterday - some comments/suggestions for you to consider:   1. You mention in your problem motivation that "No prior research has employed network analysis to tackle the problem of churn prevention". That's not strictly true :-) There's been quite a bit of recent work on the topic. So it will be nice if you can refer to some of these in your project to see if there's something you can borrow or build upon. Two examples:   [Social networks for enhanced player churn prediction in mobile free-to-play games | Applied Network Science | Full Text (springeropen.com)](https://appliednetsci.springeropen.com/articles/10.1007/s41109-022-00524-5)  [Social network analysis for customer churn prediction - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S1568494613003116)   1. For your predictive model, in addition to the centrality based network features, you can also consider other types of features generated from the network. We will briefly cover this in our lecture on predictive modeling, but one good idea here could be to generate network embedding features which are large dimensional (but non-explainable) feature sets. A popular option for this is node2vec :   [node2vec (stanford.edu)](https://snap.stanford.edu/node2vec/)   1. Lastly, you've mentioned that you're interested to explore a causal question of whether churn behavior is affected by others in the network. Maybe you can think of what would be a good way to study this question? (e.g., using a contagion model, or maybe an ERGM based model etc.). Or if you feel that you just want to focus on the predictive problem for this project, that's ok too.   Feel free to reach out to me for more discussions or clarifications!  Best,  Prasanta |

this weekend

* lit review (Jiaxin): list out report structure & estimate workload
* node2vec (Yuqi)
* run all centrality and save as csv(ziwei, Yaxin)