

Video Game Communities Knowledge Graph Analysis

Team:

Polina Haryacha

Advisor:

Amarnath Gupta

Group 6a, June 2023

Background

3.09 billion

Video Gamers Worldwide

\$245.1 billion

Gaming Market Size

\$4.4 billion

Spent on Influencer Marketing in Games

A group of young people, likely esports players, are sitting at computer desks in a gaming cafe. They are wearing red t-shirts and large gaming headsets. One person in the foreground is high-fiving another person who is smiling. The background is dimly lit with blue and purple neon lights. A white circular callout box is overlaid on the right side of the image, containing text.

The main appeal of modern video gaming is users ability to build communities around the games

Problem

◎ **Fragmented Data**

With players scattered across multiple platforms like Twitch, YouTube, Reddit, and Discord, publishers struggle to gather and integrate data from different sources.

◎ **Numbers over Influence**

Targeting is based on user reach (# of followers or impressions) without understanding of their influence in the network

◎ **Suboptimal budget allocation**

These issues lead to inefficient resource allocation, missed opportunities and suboptimal decision making



Proposed Solution

GamerGraph demystifies dynamics of gaming communities and provides video game publishers with accessible, data-driven tools to identify key players within these communities. Through the power of advanced analytics, publishers can gain valuable insights to shape their content strategy, media positioning, and partnership evaluations.



How?


1. Build a **Knowledge Graph**
2. Use **Graph Analytics** to gain insights
3. Identify **Key Players** who can effectively spread their message




Key Player Problem



The Key Player Problem (KPP) involves finding a set of nodes in a social network that either **maximally disrupt communication** among the remaining nodes (KPP-1) or are **maximally connected** to all other nodes (KPP-2).



The background of the slide is a light gray network of interconnected nodes and lines, resembling a data graph or a molecular structure. The nodes are represented by small circles, some of which are highlighted with a darker shade. The lines connecting them are thin and light gray.

Data Acquisition

Twitch



- © Twitch is a rapidly growing streaming platform owned by Amazon, with a focus on video games, attracting 140 millions active users
- © Unlike many other social media APIs, public Twitch API uncovers relationships between users in the Twitch network
- © Limitations: no historical data, no search endpoint
- © Approach: cron job for 10 days to extract all live stream data. Additional requests to get data about broadcasters, followers, and chat activity.



Steam + Steam Spy

- © Steam is a video game distribution service with 120 million monthly active users and more than 50,000 games operating live on Steam.
- © There are 2 types of reviews: user reviews and curator reviews. Steam Curators are individuals or organizations that make recommendations to help others discover interesting games in the Steam catalog
- © All games data, all curator reviews data as of February 2d, 2023
- © Steamspy API supplements Steam data by providing additional insights about games and provides additional properties for the knowledge graph

Data Quality Issues

- ◎ **Missing or incomplete data**

- ◎ **Noise**

- Bots
- Non-gaming content
- Non-English content
- “Amateur” content

- ◎ **Unstructured Data**

Typos, slang, non-english characters and emoticons in stream titles and Steam reviews

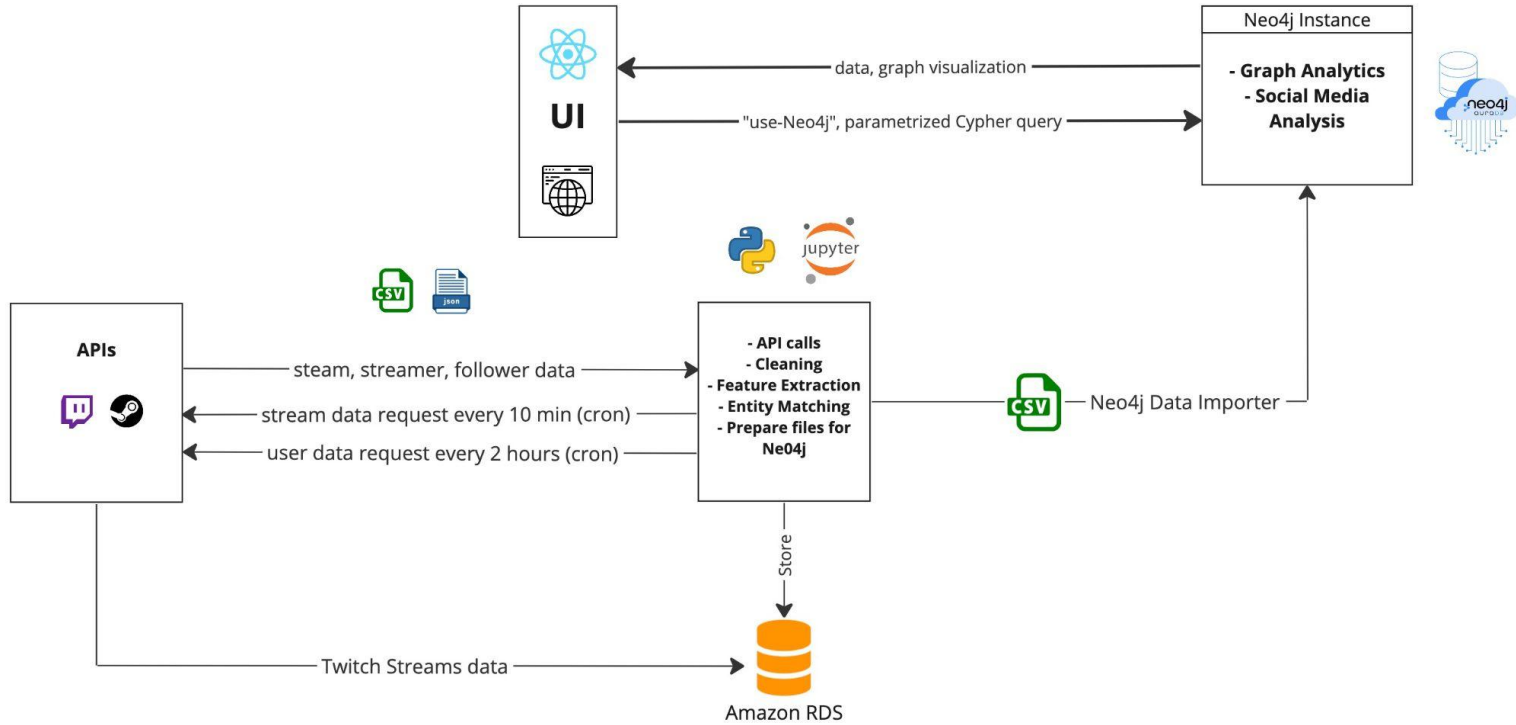
- ◎ **Duplicate Data**

Streamers, followers, chatters, and moderators overlap

Data Sizes

Source/Size	Before Cleaning	After Cleaning
Steam Games	154,750	72,234
Twitch Streams	4,974,151	163,780
Twitch Broadcasters	1,526,000	46,910
Twitch Users	6,228,377	2,609,307
Steam Reviews	942,826	757,424
Steam Curators	29,184	28,457

Data Pipeline



Feature Selection

Features: **Average Viewership, Peak Viewership**

- © For each stream, we have multiple snapshots (taken every 20 min). Each stream also includes start time timestamp. Every snapshot has a viewer count value
- © Stream-level Calculations -> Streamer-level Aggregation
- © Benefits for Advertisers: understand the general level of viewership; assess the maximum potential exposure

Feature Selection

Feature: **Sponsored Streams**

- © Identified sponsored Twitch Streams by looking at hashtags like “ad”, “sponsored” etc.
- © Tried SpaCy, Flair, OpenAI GPT-3 and GPT-J with different engines to perform Name Entity Recognition
- © Models like SpaCy trained on general news corpora, which have little relevant (gaming) content. Would require training our own model

Feature Selection

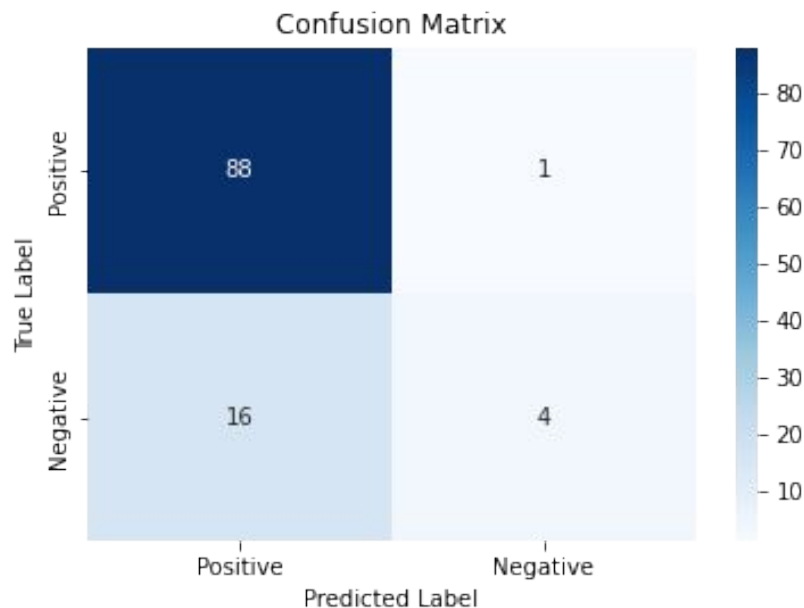
Feature: **Sponsored Streams**


Best model by training precision and recall:

GPT-J with finetuned-gpt-neox-20b.

Precision: 0.82, Recall: 0.94 (on Test data)

Insights: competitive analytics



The background of the slide is a light blue network diagram. It consists of numerous small, light blue circles (nodes) connected by thin, light blue lines (edges). The nodes are arranged in a somewhat random pattern, with some clusters and some isolated nodes. The overall effect is a complex, interconnected web of nodes and lines, suggesting a network or data structure.

Analysis & Findings

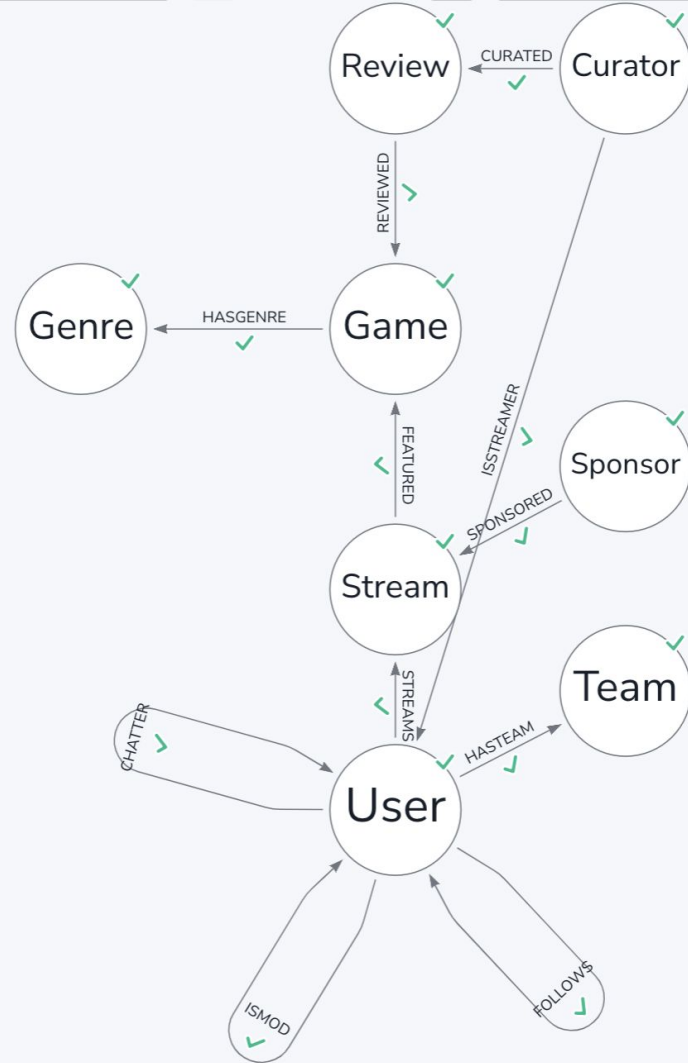
Knowledge Graph Structure

Nodes: **3,683,643**

Relationships: **5,274,191**

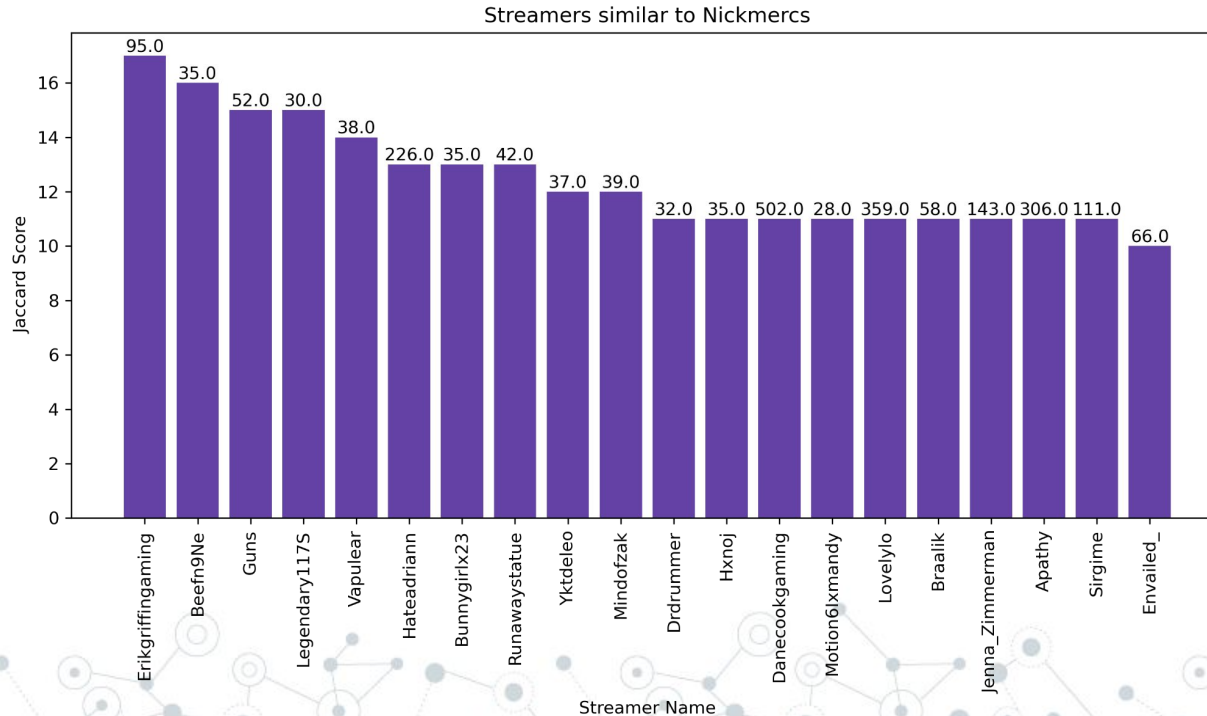


Graph Data Science as a Service



Knowledge Graph Analytics - Key Insights

Using Jaccard similarity to find “similar” streamers/curators.



Knowledge Graph Analytics - Key Insights

Identify Top 10 games popular on both platforms.

	game.game_title	steamReviewCount	twitchStreamCount	weightedAverage
0	Quell Zen	4	464	0.186963
1	Call of Duty: World at War	551	6	0.149181
2	Call of Duty: Modern Warfare 2 (2009)	371	4	0.100430
3	The Ship: Murder Party	269	14	0.077266
4	D: The Game	10	134	0.056350
5	Astral Heroes	2	113	0.045805
6	Men of War	37	87	0.044712
7	HOMEBOUND	3	64	0.026440
8	Spikit	1	59	0.023904
9	Intruder	78	2	0.021579

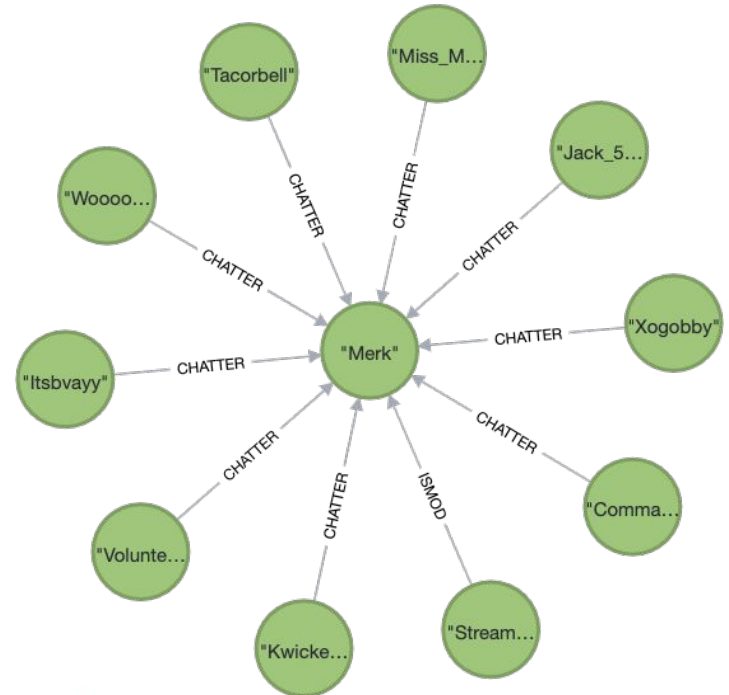
Knowledge Graph Analytics - Key Insights

Who are Top Reviewers across both platforms and what are their preferences?

	curatorName	steamFollowers	averageTwitchViews	totalfollowing	topGames	genres
0	Warhammer	253488	356.0	253844.0	[Warhammer: Vermintide 2, Total War: WARHAMMER...	[Action, Massively Multiplayer, Racing, Indie,...
1	IGN	180928	303.0	181231.0	[Warhammer 40,000: Gladius - Relics of War, Pl...	[Action, RPG, Sports, Indie, Adventure, Design...
2	Yogscast Games	141743	476.0	142219.0	[The Sinking City, Zombotron, Pillars of Etern...	[Indie, Adventure, Casual, Action, Racing, Uti...
3	Builders, managers & commanders	120378	30.0	120408.0	[Mini Metro, The Innsmouth Case, Unholy Heights]	[Adventure, Strategy, Racing, Indie, Action, S...
4	Extra Credits	103680	78.0	103758.0	[Brothers - A Tale of Two Sons, Sonic Mania, O...	[Adventure, Free to Play, Indie, Game Developm...
5	Vinesauce Vidya	94334	6665.0	100999.0	[Killing Floor 2, Planet Coaster, I Am Bread]	[Indie, RPG, Action, Early Access, Free to Pla...
6	WGN Chat	42223	39.0	42262.0	[The Disney Afternoon Collection, Stick Fight:...	[Indie, Action, Simulation, Adventure, Casual,...
7	The AngryJoeShow Army =AJSA=	38200	821.0	39021.0	[OnlyCans: Thirst Date, Far Cry 4, Divinity: D...	[Simulation, Education, Adventure, Action, Ind...
8	Giant Bomb Staff	30056	513.0	30569.0	[Marvel Heroes Omega, Divinity: Original Sin 2...	[Adventure, Indie, Strategy, Casual, RPG, Simu...
9	PsiSyndicate	28682	63.0	28745.0	[Project Zomboid, The Mortuary Assistant, Last...	[Sports, RPG, Simulation, Indie, Action, Web P...

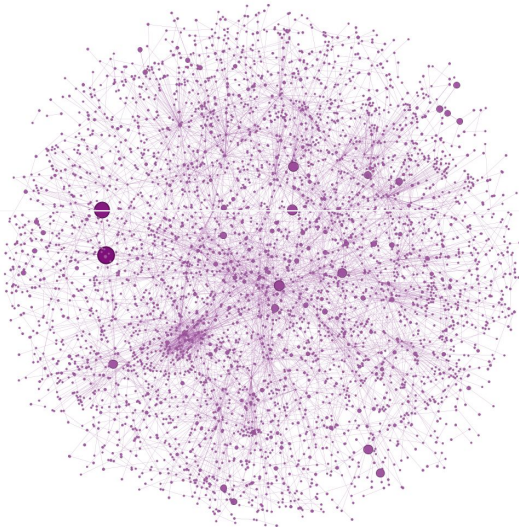
Identify Key Players in the Twitch network

- ◎ **Objective:** Identify Key Players
- ◎ **Data:** data in the database was scraped between January 31st and February 9th of 2023 (includes streamers who streamed over that period and users who chatted over that period)
- ◎ **Network Structure:** Network contains 3 types of relationships between users, namely follower, chatter, and moderator.
- ◎ **Chatters are more actively engaged users** who participate in conversations during the stream.

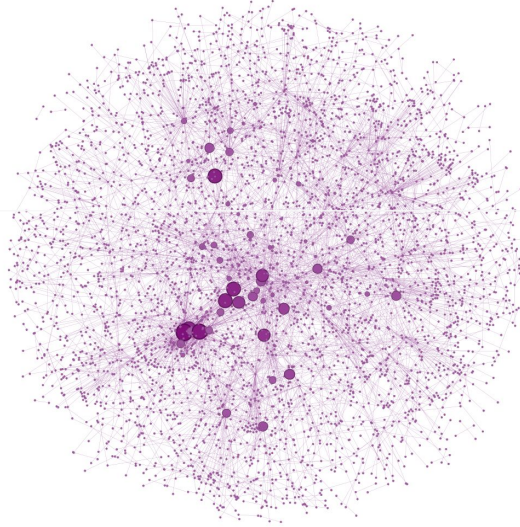


Different Centrality Measures

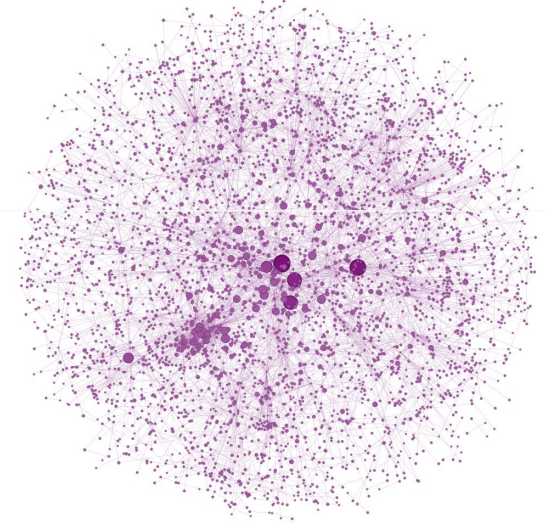
In-Degree



Betweenness



PageRank



Different Centrality Measures

In-Degree Centrality

- Indicates how many other users are chatting during the streams of the user (users who generate most engagement)
- Helps understand the level of interaction happening around the user's content.

Betweenness Centrality (Reverse Orientation)

- High betweenness centrality indicates that the user's chat messages are reaching a diverse set of streamers and viewers (most efficient “spreaders”)
- Streamer with high betweenness centrality would act as a bridge or intermediary, connecting different segments of the network.

PageRank

- PageRank can identify Twitch users who receive chat messages from other influential users.
- This suggests that the user has a significant presence and influence in the Twitch community.

Combined Centrality

- ⦿ Different centrality measures capture different aspects of node importance.
- ⦿ Combining local and global centrality
- ⦿ In-degree centrality and eigenvector centrality are combined to calculate the combined score.

$$\text{CombinedScore} = (0.5 * \text{degreescore}) + (0.5 * \text{eigenvectorscore})$$

- ⦿ Degree centrality represents popularity and reach, while eigenvector centrality considers the influence of a streamer's connections.
- ⦿ **The combined score aims to provide a balanced representation of a streamer's overall influence, considering both their own connections and the quality of those connections.**

Evaluation Challenges

Gaining accurate ground truth for the influence of nodes in the Twitch network is a challenging task.

Potential Approaches:

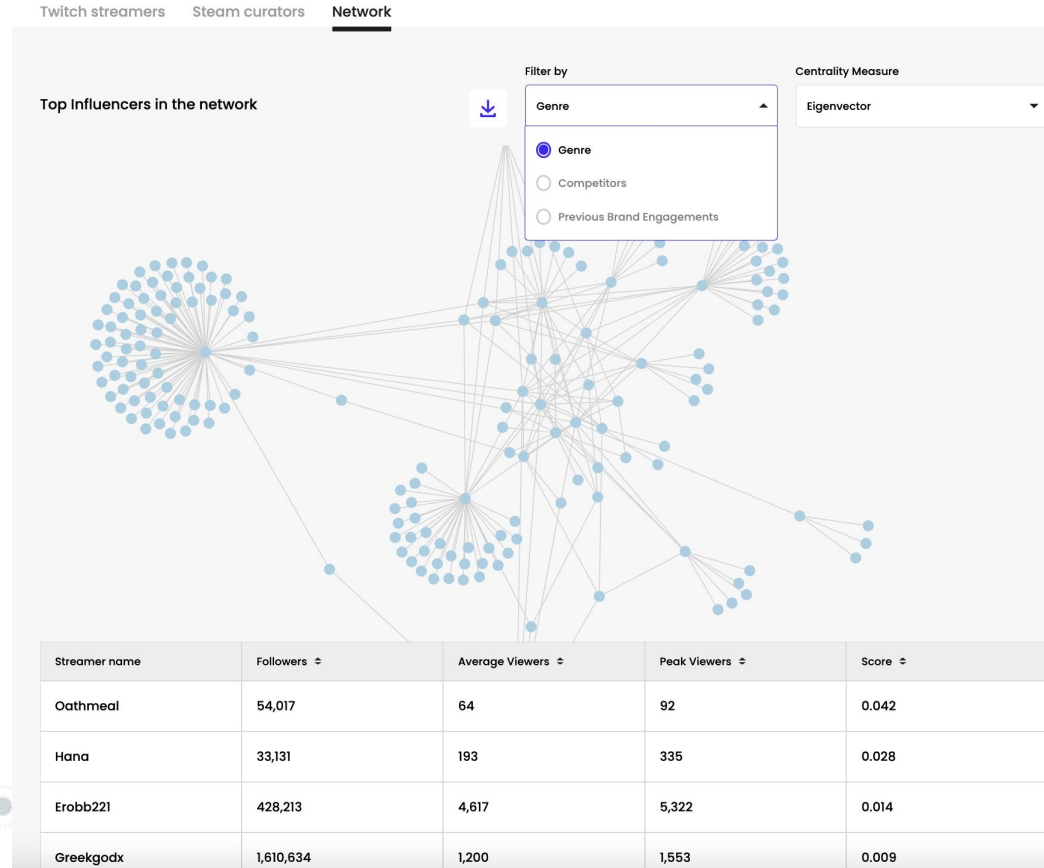
- ⦿ Analyzing historical data
- ⦿ Conducting surveys and interviews with Twitch users
- ⦿ Expert evaluation and industry knowledge
- ⦿ Simulate a theoretical diffusion model (simulation of the SIR diffusion model with predicted node as the seed of diffusion)

The background of the slide is a light blue-grey color with a repeating pattern of interconnected nodes and lines, resembling a network or molecular structure. The nodes are small circles, some solid and some hollow, connected by thin lines.

Reporting

Reporting Dashboard

- ◎ Layout design implemented using React with TypeScript
- ◎ Use-neo4j Hooks to send parameterized Cypher queries to the Neo4j database
- ◎ Recharts to display data visually
- ◎ [Force-graph](#) to visualize graph
- ◎ Vercel to test and preview



Case Study



Ann
Marketing Analyst
@ TinyBuild Games

1. Beta Phase

Goal: gather feedback, test the game, and build a dedicated community of early adopters.

Target: smaller Twitch Streamers who play games similar to upcoming title, who are RPG fans and have a highly engaged audience (many chatters)

2. Game Launch

Goal: big budget, strong impact.

Target: Twitch streamers and Steam curators with high viewership and high PageRank.

3. Similar Streamers

Goal: Ann has been following streamer named “Nickmercs”, his streams consistently attract high engagement for similar titles. She wants to find a group of similar streamers and recruit them for the Brand Ambassador program.

Target: find streamers similar to Nickmercs based on shared properties (games they play) and shared followers.

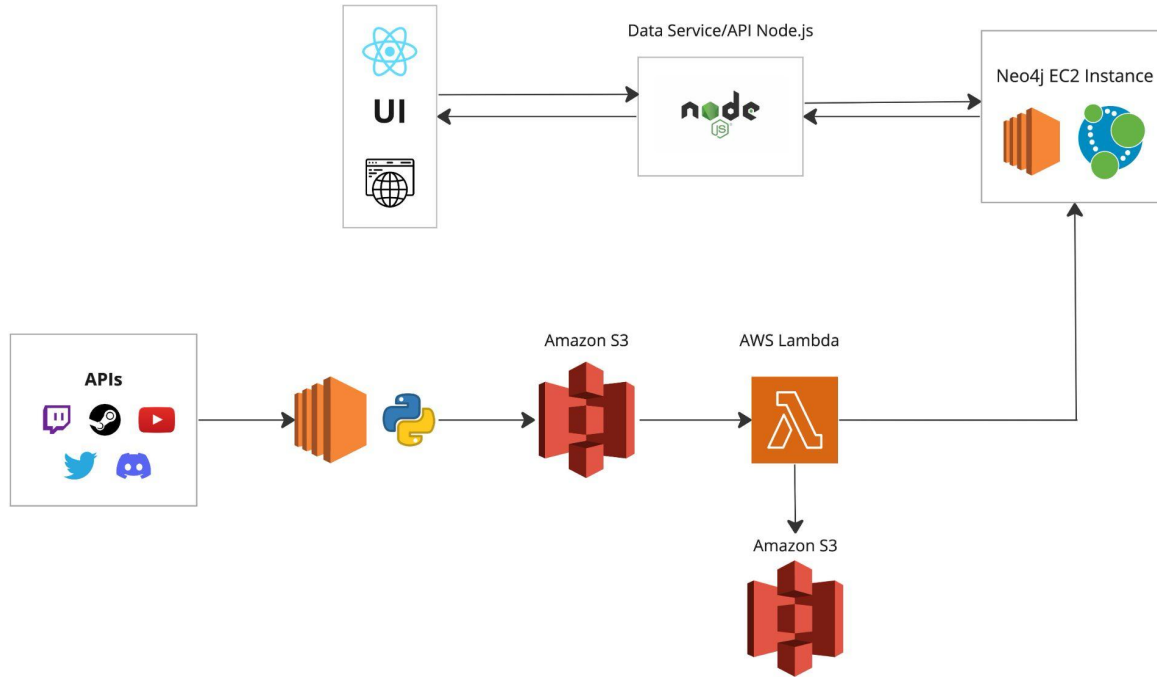
The background of the slide is a light blue-grey color with a repeating pattern of a network diagram. The diagram consists of numerous small circles (nodes) connected by thin, light grey lines (edges). Some nodes are highlighted with a darker grey or blue color, and some are enclosed in a dashed circle. The overall effect is a dense, interconnected web of nodes and lines.

Demo

The background of the slide is a light blue network diagram. It consists of numerous small, light blue circles (nodes) connected by thin, light blue lines (edges). Some nodes are highlighted with a darker blue color, and some are enclosed in a dashed circle. The overall pattern is a complex, interconnected web of nodes and lines, suggesting a network or data structure.

Solution Architecture

Proposed Architecture (Scalability Plan)



Challenges and Next Steps

Discoveries and Challenges:

- ⦿ Add more data sources: YouTube, Discord, Reddit etc.
- ⦿ Scalable and automated backend system. The product is not yet ready for deployment, and a scalability plan needs to be implemented to handle the high volume of data
- ⦿ Major challenge in validating the results of key player identification
- ⦿ Combined Centrality measure - there are other ways to combine centralities

Acknowledgments



Amarnath Gupta



Ilkay Altintas



Thanks!

Any questions?

