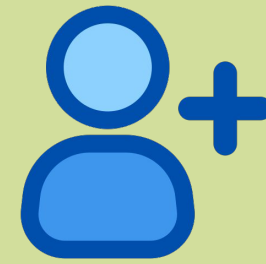# The Social Connection

Analysis of Social Connections from Google+

Alex Oshima | Dominic Orlando | Ishita Chawla | Yumeng Tang | Jasmine Le

Motivation

How do friend groups form online?

- Understanding how social circle forms can reveal real-life and online relationships

Who has the most connections?

- In every friend circle, some people are more connected than others
- We can find the connectors and see how the info spreads

Can AI Predict who you'll be friends with?

- We are analyzing what influences social group formation (how social media platforms recommend connections)

Objective

- Analyze Ego (User) Networks to understand how people group into social circles

- Identify the key features that influence these social connections

- Use machine learning (recommender system) to predict social circles/network structures
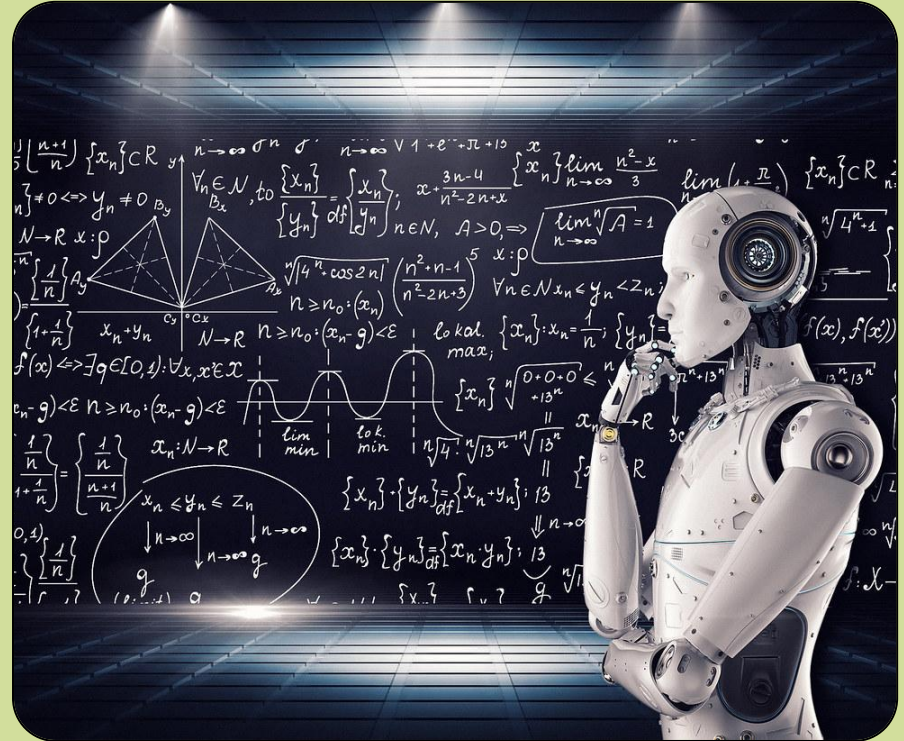
Methodology

1. Data Exploration and Pre-Processing

2. Model Training and Data Analysis

3. Evaluation and Comparison of Results using Visual Insights

Dataset

## Learning Social Circles in Networks

- Social connection centered around an individual (Ego Network)
- Shows who is connected to whom
- **Social Circles** → Groups of friends a user belongs to
- **User Features** → Includes details like workplace, location, education, etc

- Helps us understand how people form social circles online
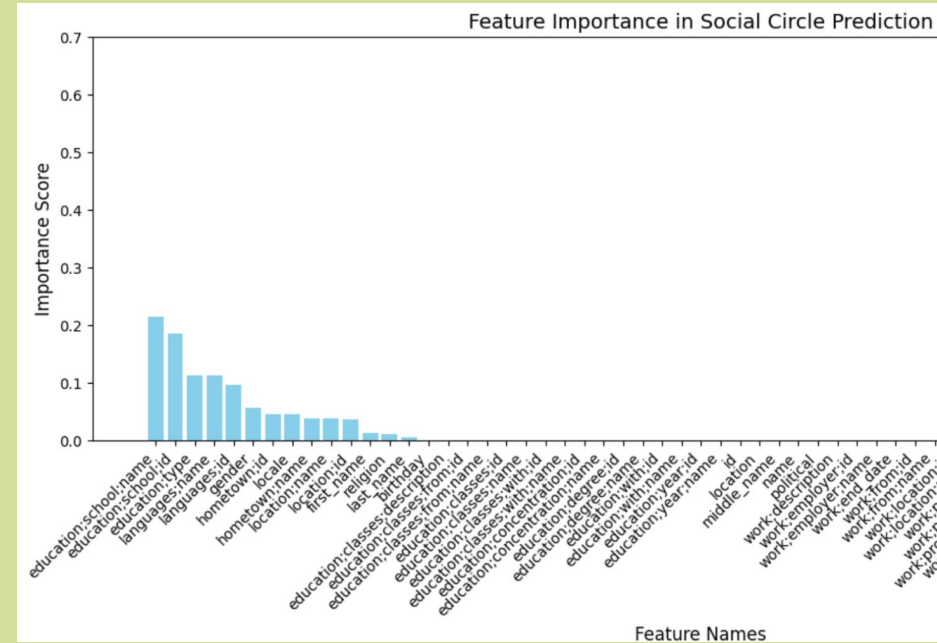- Used for predicting relationships & recommending connections

Overall Feature Importance vs. Individual Circles

Overall Feature Importance Across All Social Circles
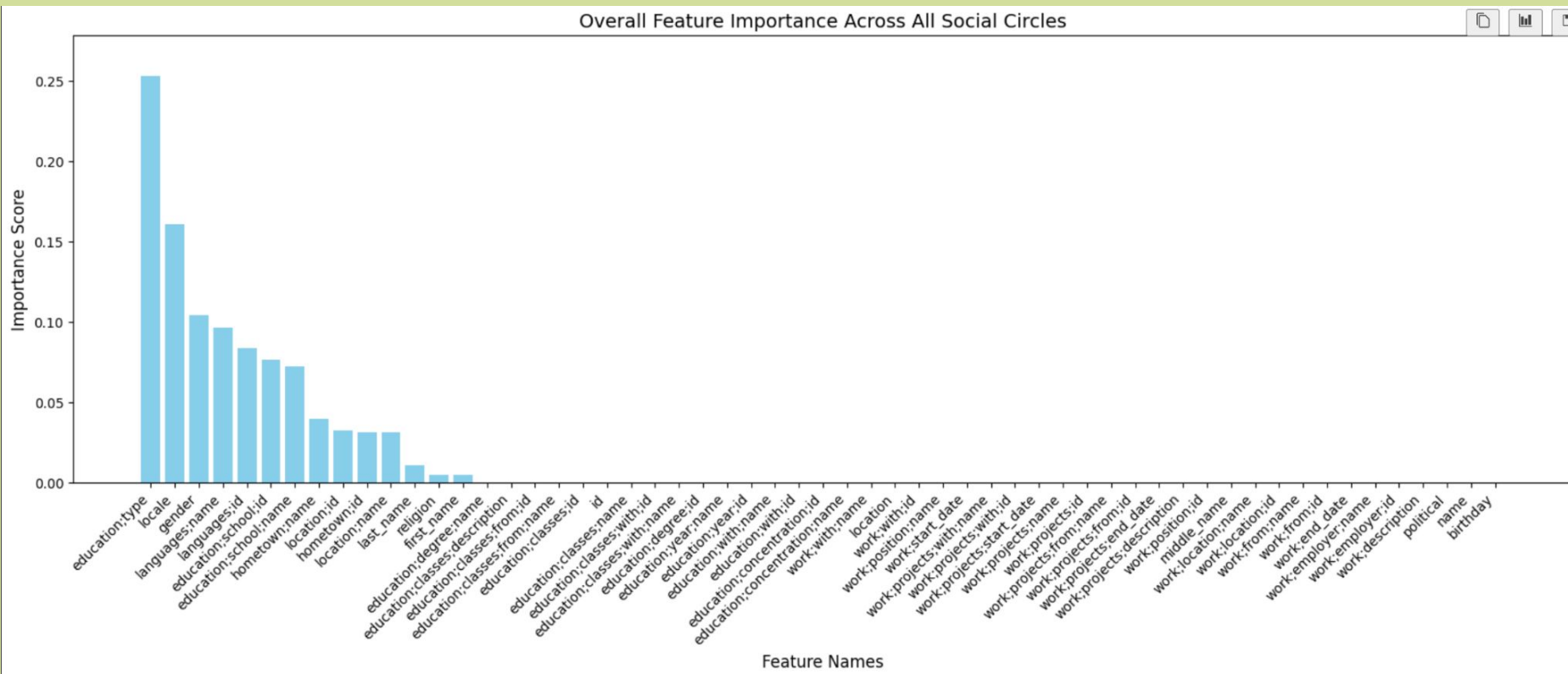
Overall Feature Importance vs. Individual Circles
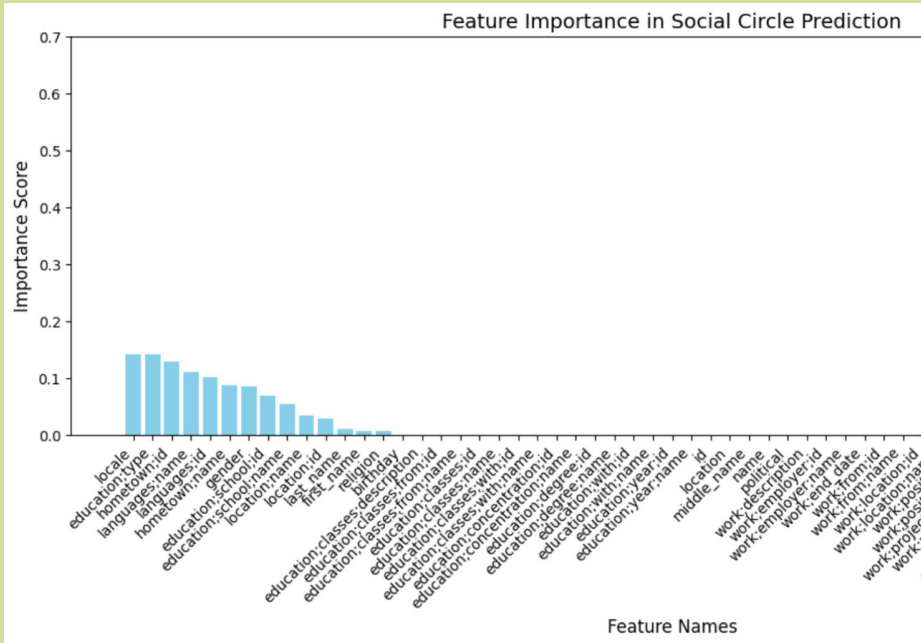
Circle 29

Circle 22

Ego 345's Full Social Circle
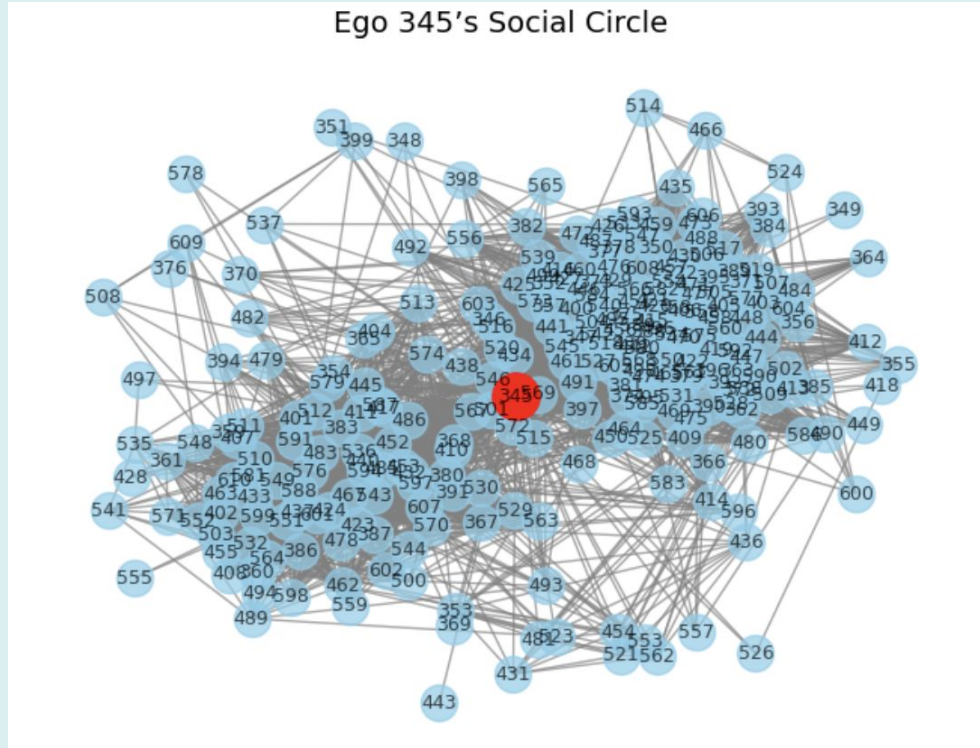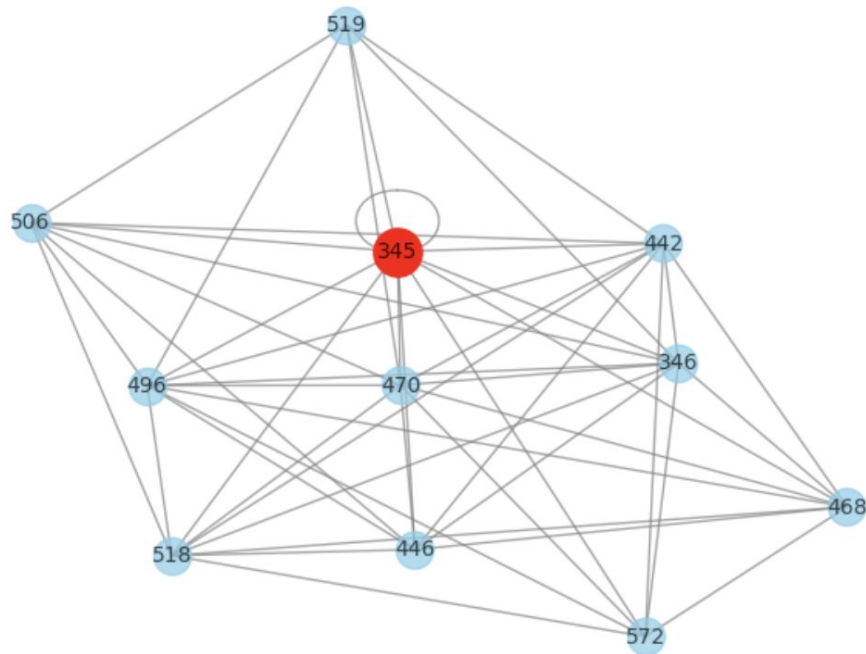

Ego 345's Social Circle

Ego 345 's Close Social Network

Ego 345's Social Circle (First 10 Direct Friends)

Boosting

See what we can get!

Mask one connection and train the model to recover if there's a connection or not
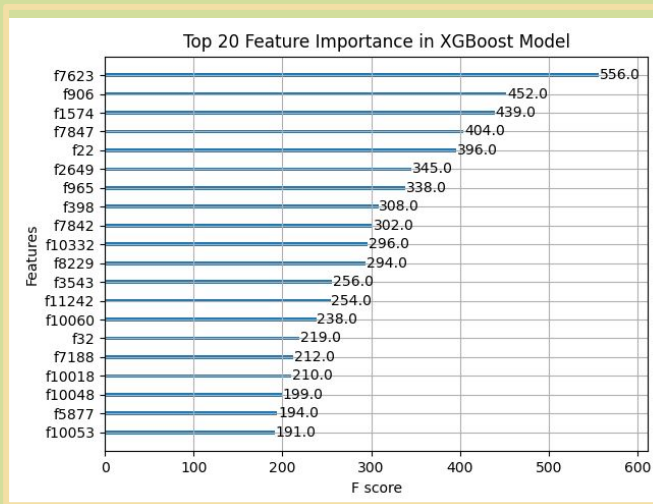
Only use people connection data

Use as recommender system by recommending errors :D
(If people aren't connected but model thinks they should be, then recommend it)

Accuracy: 0.9615
Precision: 0.9479 Recall: 0.9768

Top 20 Feature Importance in XGBoost Model

| Features	F score
f7623	556.0
f906	452.0
f1574	439.0
f7847	404.0
f22	396.0
f2649	345.0
f965	338.0
f398	308.0
f7842	302.0
f10332	296.0
f8229	294.0
f3543	256.0
f11242	254.0
f10060	238.0
f32	219.0
f7188	212.0
f10018	210.0
f10048	199.0
f5877	194.0
f10053	191.0

Average feature importance: 13.34
Features are people, so shows how impactful being connected to someone is

Similarity-based Link Prediction

This method predicts connections by comparing shared neighbors between two nodes. If two people have many mutual friends, they are more likely to be connected.

In this dataset, we compute Jaccard Similarity for each node pair and predict a link when they share a significant fraction of common neighbors.

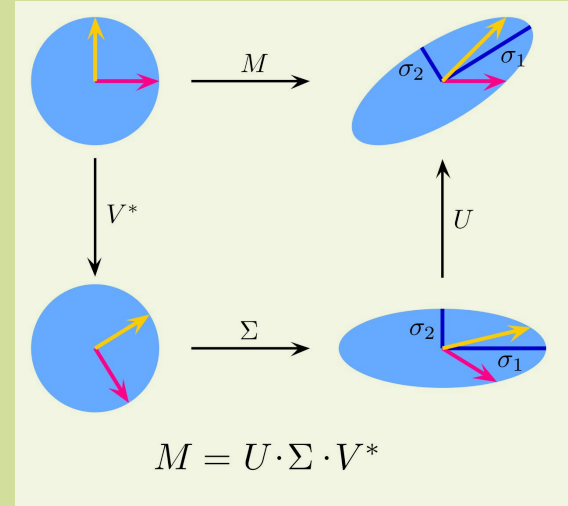$$Jaccard = \frac{intersection(A, B)}{union(A, B)}$$
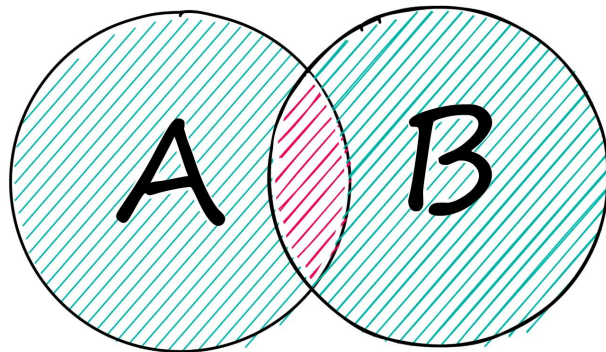
Accuracy: 0.9932

Singular Value Decomposition(SVD)

SVD decomposes the network structure into lower-dimensional representations, capturing hidden relationships between nodes.

Here, we apply SVD to the adjacency matrix, using the learned node embeddings to estimate the likelihood of a missing link.
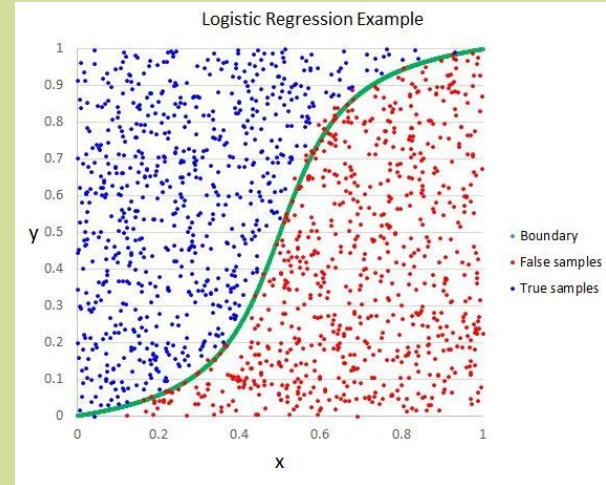

$$M = U \cdot \Sigma \cdot V^*$$

Accuracy: 0.9865

Logistic Regression

We can model the connections between individuals as a linear combination of parameters

We can utilize the most impactful parameters to train our model, resulting in similar accuracy with quicker computation

Logistic Regression Example

- Boundary
- False samples
- True samples

Accuracy: 0.9996

Comparison

Model	Accuracy	Running time
Boosting	0.9615	6 hours
Jaccard Similarity	0.9932	2.3 seconds
SVD	0.9865	189.7 seconds
Logistic Regression	0.9996	187.3 seconds