Homophily - A Driving Factor for Hate Speech on Twitter

International Conference on Complex Networks, 2021

Background and Motivation

1. Homophily is defined as the tendency of like-minded (similar) people to connect/befriend (familiar)
2. Homophily structures a user's ego network on social networks
3. Homophily plays a significant role in information diffusion and dissemination
4. Homophily is a driver factor in product adoption, online guild formation, sustenance and community formation
5. But homophily has not been studied in generation of hate speech

Our Approach

The proposed approach has three main components:

1. **Define a novel metric for familiarity computation**:
   - We use a VGAE to encode a user's position in the social network
   - We hypothesize that these graph encodings can capture a user's position as well as their society from a network's perspective

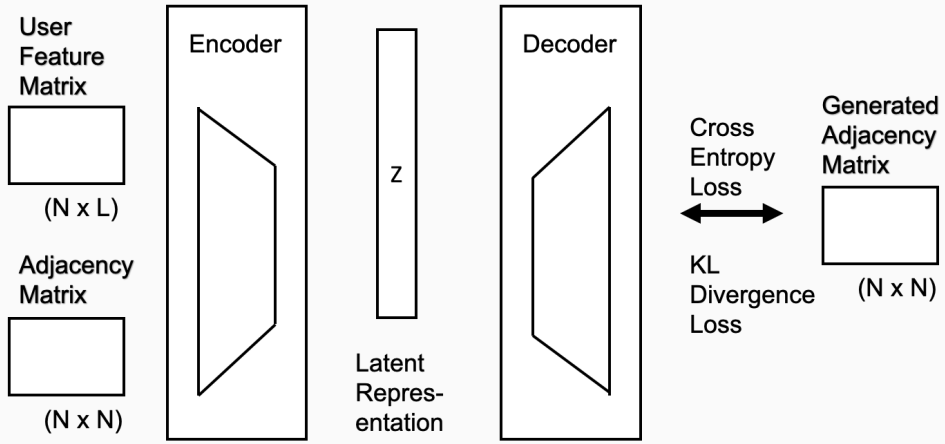

Figure 1: Variational Graph Auto Encoder

Our Approach (Cont'd)

2. **Utilizing the novel metrics in showing homophily in hate speech**:

Let $\bar{u_1}$ and $\bar{u_2}$ represent graph embedding of the users u_1 and u_2 respectively.

$$CosineSimiliarity(u1, u2) = \frac{(\bar{u_1} \cdot (\bar{u_2})}{||\bar{u_1}|| \cdot ||\bar{u_2}||} \tag{1}$$

3. **Detecting hateful forms on social media platforms**:

- We use latent topic modelling to detect multiple hateful forms present in hate speech
- We hypothesize that individual hateful forms, differing in nature, might exhibit varied homophilic behaviours

Experiments

- Dataset
  - We use hate speech dataset provided by ""Like Sheep Among Wolves": Characterizing Hateful Users on Twitter"
  - It has 200 recent tweets of 100,386 users along with retweet induced graph
  - 4,972 users are labelled as hateful. We pick these users along with some more users
  - Manually annotate these users', 30,720 tweets as hateful or not.
- Research Questions
  - **RQ1**: Is homophily exhibited by the users generating hateful content?
  - **RQ2**: How effective is the newly proposed familiarity metric in comparison to the existing?
  - **RQ3**: Is homophily more prominent for particular forms of hate?

Experimental Results

- To answer **RQ1** and **RQ2**, we plot similarity against familiarity for the three types of familiarity metrics in Figure 3. As the hatefulness increases, homophily also increases
- We compare the Pearson correlation coefficient between similarity and familiarity.
- Our proposed metrics result in highest coefficient of 0.6, while mutual friend based results in least of 0.2

Experimental Results (Cont'd)

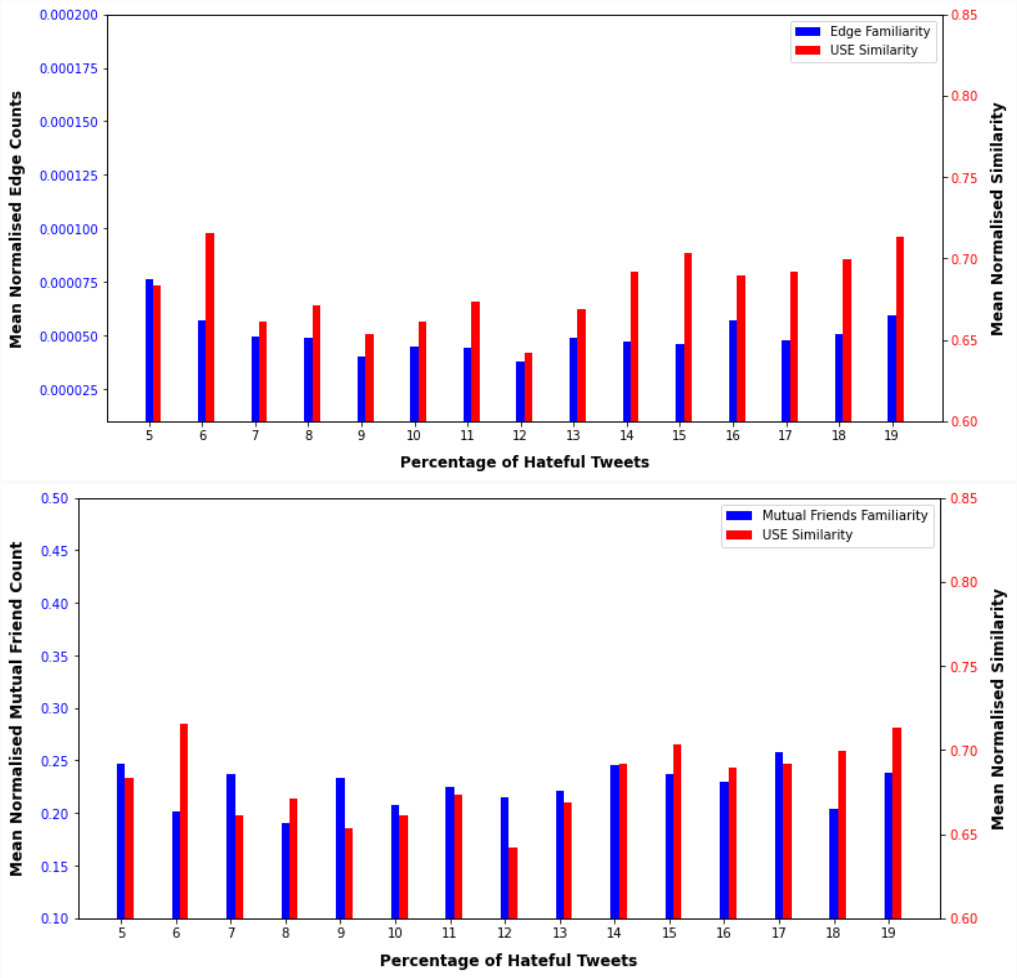

Figure 2: Variation in similarity and familiarity as hatefulness increases

Experimental Results (Cont'd)

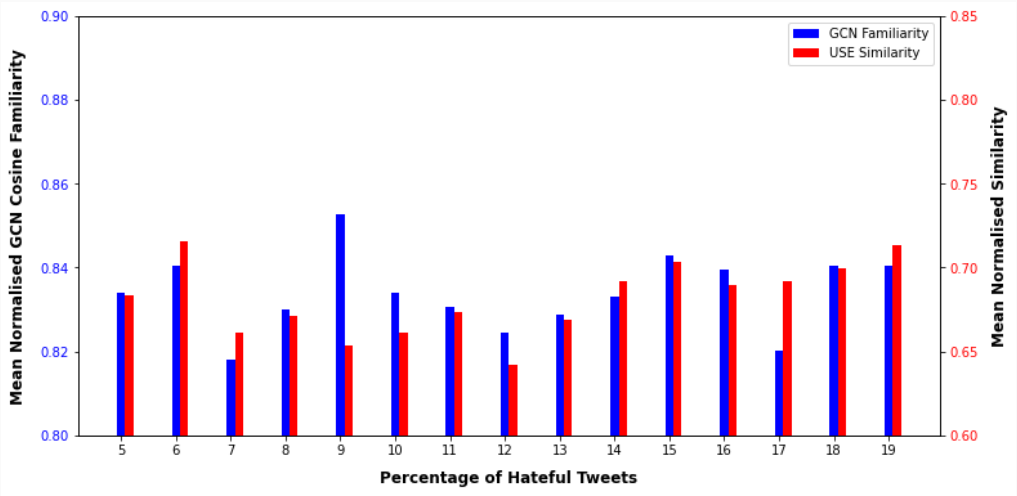

Figure 3: Variation in similarity and familiarity as hatefulness increases

- To answer RQ3, we create a user base for each hateful form (topic). We pick users whose affinity score is above a certain threshold.
- We also rank the different hashtags used by users by frequency. This is shown in Table 1.

Experimental Results (Cont'd)

Table 1: Top Hashtags for the Hateful Topics

Topic	Hashtags
0	#maga, #trump, #realdonaldtrump, #trumptrain
1	#impeachtrump, #trump, #trumprussia, #jfkfiles
2	#bitch, #metoo, #harvey, #lockherup
3	#gobills, #pelicans, #mlscupplayoffs
4	#london, #fakenews, #cancer, #queen
5	#tormentedkashmir, #kashmirsuffering, #pakistan
6	#brexit, #crime, #terrorism, #illegal
7	#nigga, #bitch, #bitches, #somalia, #nigger

- We observe that topics 2, 5 and 7 exhibit stronger homophily, as compared to others for both the communities.
- These topics can be broadly categorized into hate manifesting as sexism, nationalism and racism.

Experimental Results (Cont'd)

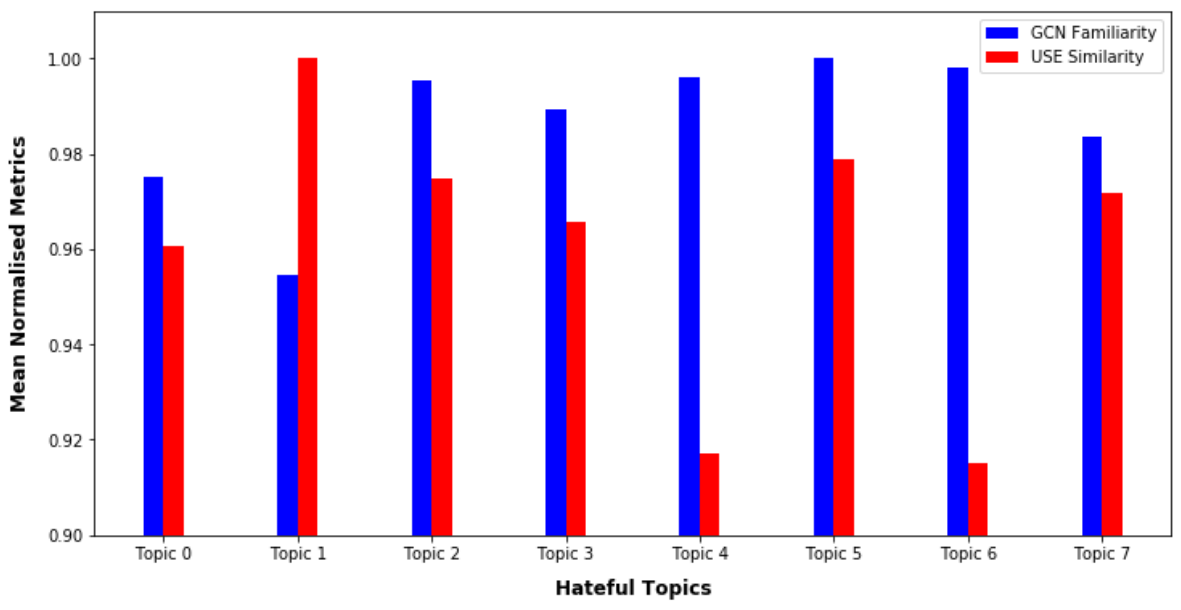

Figure 4: Homophily for the different hateful forms

Summary

1. We propose a novel way to compute familiarity using graph embedding technique.
2. We show homophily in hate speech on a dataset from Twitter.
3. We empirically demonstrate the effectiveness of the newly proposed metrics in establish familiarity against the existing metrics, using homophily as the benchmark of comparison.
4. We do a deep dive analysis of variations of homophily in different forms of hate.