HY-484 Final Project Presentation

CLUSTERING AND MISINFORMATION

Evangelos Kalimeris

UNIVERSITY OF CRETE

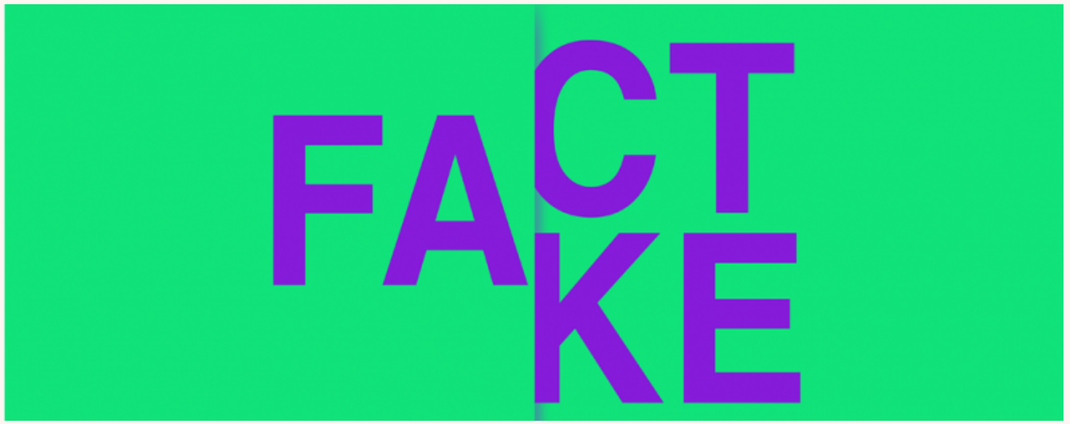
MOTIVATION

MISINFORMATION

- Is a chronic problem, enhanced by online communities
- The claims range from hilarious to dangerous
- Most people don’t have scientific background
- Online social media services try to detect false claims
- Is this enough ?

HOW CAN WE HELP ?

- Social media services passively try to detect fake news and label/remove them
- Is this efficient or enough ?
- Is this ethical ?
- How could we actively change online societies behaviour ?
- Many individuals fall victims of spreading misinformation and would like to avoid it
- By changing the behaviour of some, we affect more by possible cascading



HYPOTHESIS

THE EMPEROR'S DILEMMA

- A hilarious or extremist notion, certainly not backed by any scientific community, could only be spread by individuals belonging to small social circles.

- We want to find out if this is true on online social media

- In favour of this claim, would be any of the following findings:
  - Fake news spreaders having less total friends than real news spreaders
  - Fake news spreaders having more friends who are fake news spreaders than not
  - Fake news spreaders belonging to smaller interconnected clusters instead of larger Fully connected components

METHODOLOGY AND FINDINGS

DATA

- Two datasets were used from BuzzFeed and POLITIFACT news sites respectively
- Subjects were Twitter users who twitted real and fake news being fact checked and available on both sites
- 15257 total users from BuzzFeed
- 23865 users from POLITIFACT

FIRST TASK

- Python scripts -

- In both datasets users posting real or fake news were mostly discrete
- 48% posted only real, 48.5% posted only fake news on BuzzFeed
- 18.6% posted only real, 79% posted only fake news on POLITIFACT
- Only 3.5% and 2.4% posted both real and fake news, and was dismissed as a group from this study, as insignificant
- This clear segregation helped further analyses

HOMOPHILY

- **BuzzFeed:**

$$A = p^2 + q^2 = 7316^2 + 7406^2 = 108372692$$
$$B = 2 \cdot p \cdot q = 2 \cdot 7316 \cdot 7406 = 108364592$$

- **POLITIFACT:**

$$A = p^2 + q^2 = 4437^2 + 19428^2 = 397154153$$
$$B = 2 \cdot p \cdot q = 2 \cdot 4437 \cdot 19428 = 172404072$$

- **As we can see A > B in both cases, so both networks are homophilic**

SECOND TASK

- Python scripts -

- Twitting amount of real news posters was marginally greater
  - 1.41 versus 1.37 tweets per real versus false poster respectively on BuzzFeed
  - 1.51 versus 1.27 tweets per real versus false poster respectively on POLITIFACT
- Re-twitting amount had insignificant differences among groups, all of them having close to 1.1 retweets per user
- Real posters appear slightly more active
- Both groups appear as persistent

THIRD TASK

- Python scripts -

- **Social characteristics - Twitter ‘follow’ relationships**

- **BuzzFeed:**
  - **Real posters had 40.6 followers and followed 40.9 others on average**
  - **Fake posters had 42.6 followers and followed 41.9 others on average**

- **POLITIFACT:**
  - **Real posters had 24.8 followers and followed 25.3 others on average**
  - **Fake posters had 24 followers and followed 23.6 others on average**

- **The results were found comparable and not indicative of the hypothesis**

FOURTH TASK

- Python scripts -

- **Diversity measurement - Do real spreaders have more friends of both groups than fake spreaders?**
- **Buzzfeed:**
  - Both real and fake spreaders had on average 19-21 connections to both groups
- **POLITIFACT**
  - All group combinations besides fake → fake had on average 4.5-4.9 connections
  - Fake → fake connections were 4 times higher, at 18.4 on average, but remember that fake posters are 4 times more on this dataset
- **The results again are not supportive of the hypothesis**

FIFTH TASK

- Gephi analysis -

- Each dataset was split into four, with each subgraph having real→real, real→fake, fake→real, fake→fake edges respectively
- Strong homophilic relationships were observed in both networks, with real→real and fake→fake subgraphs being denser, higher clustering coefficient and smaller total amount of strongly connected components [table1 / table2]
- Again both real as well as fake spreaders had similar metrics

SIXTH TASK

- Gephi analysis -

- Each dataset was split into two subgraphs containing real→all and fake→all connections respectively
- Fake news tweeters had a slightly higher clustering coefficient and less strongly connected components [table3 / table4]
- This small difference is actually opposing our hypothesis, as fake news spreaders appeared overall more connected than real news spreaders

CONCLUSION

- **Our original hypothesis was not confirmed, at least with the two datasets studied here**
- **Most metrics were found marginally different or conflicting**

FUTURE WORK

- Same analysis on more data
- The small amount of users that posted both real and fake news could prove of interest
  - Possibly easier to change their attitude
  - Possibly more integrated, with higher social influence
- Study of qualitative differences of real versus false spreaders posts

THANK YOU FOR YOUR ATTENTION