Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram

What is Instagram?

- A media sharing platform for users to reflect their interests
- Used by marketers and brands to reach their potential audience for ads.
- The number of likes on posts serves as a proxy for social reputation of these users.
- Most cases, social media influencers with an extensive reach are compensated by marketers to promote products.
- To get more benefits, users often artificially increase the popularity and engagement on their content in several ways.
    - Leverage bots
    - Purchase social metrics such as – likes, followers, and shares from black market services
    - Become part of collusion networks which can be used to trade inorganic engagement.

Goal of the study

- Artificial bolstering of popularity can cause brands to lose money, advertisers to not reach the relevant audience, and recommender algorithms to give poor suggestions.
- The study proposes that the true reach / social-worth of the user should be determined by canceling out the effect of fake engagement which the user receives, and should largely depend only on the organic engagement.
    - Given a liker *L*, who likes a specific post *p* of a poster *S* – Find out the features of *L*, *p* and *S*, to determine the probability of liker *L* genuinely liking a post *p*.
    - Contribution:
        - Characterizing Fake and Organic Likes
        - Automatic Detection of Fake Likes

Data

- Fake Like Instances:
  - Sources- paid web services, trading platforms, bots
  - IG allows videos and maintains its view and like count
- Random Like Instances:
  - Since Instagram does not provide a direct way to sample random users/posts, the authors obtained a seed set of Instagram users, and extracted their follower and followee connections in a breadth-first-search manner.

Table 1: Dataset of Instagram like instances. We also collect meta-information of likers, posts and posters.

	#likes	#posts (p)	#Likers ($\mathcal{L}$)	#Posters ($\mathcal{S}$)
FakeLike_data	16,448	9,932	9,301	7,822
RandLike_data	134,669	1,717	47,233	738

Analysis

- It is virtually impossible to know why a user might like a post, however, it is possible to understand how the user could have come across the post
- Definition: *Given a poster S whose post p has been liked by liker L, we define a like instance as the tuple (L, p, S)*
- Do not assume that if a single like generated by a liker is fake, then all her other likes are also fake
- Network Effects:
  - H1.1: *A liker L is more likely to genuinely like S's post if L is a follower of S.*
  - H1.2: *A liker L is more likely to genuinely like S's post if L is a follower of S's followers.*
  - **Genuine likers do indeed like their followee's post more than fake likers do.**
    - In case of fake engagements, only 16.8% of likers of a post are followers of the poster, as compared to a much higher fraction of 39.1% likers being followers in case of random like engagements
    - In a two-hop network, only 2.8% of likers of a post in case of fake likes are follower-of-follower of the poster, as compared to 42.4% in case of random like engagement

Analysis (cont)

- Interest Overlap:
  - H2: *A user L will have a higher chance of genuinely liking S's post if L and S share interests.*
  - Interest Profile:
    - Topic Extraction
      - Infer topics from textual sources such as bio and post captions using Wikification
      - Used Densecap captioning to obtain meaningful captions from images; Wikification to extract fine-grained topics
    - Topic Matching
      - Word2vec similarities between two tuples of interest
  - **60% of fake likers have an affinity value of 0.475, as compared to 0.58 affinity for same proportion of random set of likers**

Analysis (cont)

- Liking Frequency:
  - H3: *A liker $\mathcal{L}$ will genuinely like more than one post of the poster $\mathcal{S}$.*
  - **90% posters with fake likes get 7% repeated likers on their posts, as compared to the same fraction of posters with genuine likes getting 42% repeated likes.**
- Influential Poster:
  - H4: *A user $\mathcal{L}$ will have a higher chance of genuinely liking $\mathcal{S}$'s photo if $\mathcal{S}$ is an 'influential' user or a celebrity.*
  - In the dataset, only **1.9% users were celebrities who got fake likes, as compared to 7.5% celebrities who got genuine likes**, indicating that celebrities are more likely to attract a higher number of likes.
- Link Farming Hashtags to get Fake Likes:
  - H5.1: *A user $\mathcal{S}$ is more likely to attract fake likes if she uses link farming hashtags in her posts.*
  - Curate a list of 112 such hashtags
  - **20.8% posts with fake likes have at least one link farming hashtag as compared to 1.8% posts with random likes.**

Analysis (cont)

- Topical Hashtags:
  - H5.2: *A user S with genuine likes will have topical hashtags in their posts.*
  - A two-step process to detect topical hashtags.
    i. Filter out all link farming hashtags as well as popular non-topical hashtags.
    ii. Segment these hashtags and use Wikifier to see what proportion of hashtags pertain to a topic.
  - **Fake like instances tend to have a lower proportion of topical hashtags.**
- Detecting Fake Likes:
  - Building a Classification Model:
    i. Trained a supervised model on the above features with fake likes as the positive class.
    ii. Experimented with different classification algorithms viz. Logistic Regression, Random Forest, SVM (RBF kernel), AdaBoost (with Random Forest as base initiator), and XGBoost.
    iii. Used a simple feed-forward neural network – Multi-Layer Perceptron (MLP) for best performance.
    iv. In all the experiments they performed 10-fold cross-validation, using 80% of the dataset as training data and 20% for validation.

Results

- **Baseline:** Used a method to detect fake likers on Facebook [2]. A supervised method for the detection of fake likes based on profile (length of biography, lifespan of account, number of bidirectional connections), posting activity (average number and maximum posts per day, total posts and skewness of posting), page liking (category entropy of pages liked, proportion of verified pages) and social attention (average number of likes and comments received) of the liker.

- The baseline model gives a precision of 61%, a recall of 69%. Compared to the baseline model, their model obtains an 83% precision and 81% recall to detect fake like instances.

Conclusion

- Built on existing content-based techniques of fraud detection on OSNs by incorporating factors that motivate liking on Instagram, such as liker-poster interest overlap.
- Also account for Instagram's visual aspect by examining the contents of images. Their automated method is able to detect fake likes with 83% precision (22% increase on the baseline).