Popularity on Social Media: Exploration and Analysis of General Causal Factors

Achieving focal positions and popularity in social media networks is not a random process. If you want to be popular, you obviously do well to post interesting content and post it frequently rather than lurk and keep silent. That is, there are some particular online behaviours you should adopt if you want to be popular and others that you should avoid. Our project is based around identifying these factors responsible for popularity on different social media platforms.

Our chief research questions are therefore:

1. What characteristics of online behaviour correlate with online popularity?
2. Are these characteristics uniform or different across different social media platforms? For instance, if e.g. frequency of posting is correlated with popularity on Twitter, does the same go for popularity on Youtube?
3. What is the causal connection between these factors and online popularity? Are all of the identified factors causally responsible for popularity or are some of them merely correlated? If the former, are they independent causal factors or not?

The implementation of the project is the following.

Step I: Data Collection

Look at data availability via API and web scraping from different social media platforms: Youtube, TikTok, LinkedIn, Twitter, BlueSky, Mastodon, etc.

Identify the type of social media that has most useful and accessible data – text-, video-, or image-based (suspicion is that it's gonna be text-based).

Obtain data via APIs or web scraping. Clean it.

Identify common attributes across datasets so we can meaningfully compare results we obtain from them. In other words, if 'frequency of posting' is a parameter of interest, it should be present in data sets from all platforms we are dealing with. Conversely, if some dataset has a parameter that is not present across platforms (e.g., 'employed/unemployed' on LinkedIn), it is a sign not to use the dataset since we cannot meaningfully compare it with others for the purposes of our inquiry.

Tools: Python, `pandas` for cleaning and structuring data, `pyplot` for exploration, `BeautifulSoup` / `requests` for scraping / APIs

Step II: Define a Virality Index

This is an open question: when do we want to say that a person is popular on social media? It cannot simply be the amount of followers (e.g., a celebrity can have mass online following while posting literally

nothing) or the amount of likes (a person whose posts have on avg a lot of likes but are infrequent may be less influential than a person with less likes on avg but more frequent posting). Our virality index will probably be some combination of several such indicators. This is something to read/think about.

Step III: Train and Evaluate a Model Predicting the Index on Data Sets From Step I

Target variable is Virality Index from Step II. Predictor variables are those obtained from datasets in Step I. Obviously, parameters that the virality index depends on cannot be included in the predictor variables since they determine the index uniquely.

The model must be able to predict the virality index on the basis of the predictor parameters. Since the index is continuous, we are probably looking at some regression model (?).

Tools: Python, `sklearn` / `Keras` / `PyTorch` depending on the team's wishes/expertise and required functionality (to be specified along the way).

Step IV: Extract Features of Importance

That is, identify attributes that a good model obtained after tweaking and evaluating in Step III considers most important for prediction of the virality score. How to do this is an open question – that being said, some models in Python have a built-in method with this functionality. For instance, `RandomForestClassifier.feature_importances_` from `sklearn`.

$\Longrightarrow$ **I think that executing Steps I-IV for several social media platforms, we will already have enough results to explore and present. If it so happens that we have more time, I propose we do the following steps.**

Step V: Analysing Causal Connections via Dynamic Network Models

Now, we enter a somewhat creative territory where we can play around. After Steps I-IV, we have an indicator of popularity (virality index) and a set of parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ that we know correlate with user i's popularity V_i. That is interesting but incomplete knowledge – in particular, because we do not know *if* $\boldsymbol{\theta}$ cause higher V_i, *why* they cause it and *how*.

One way to test the hypotheses we come up with in response to above questions is to construct and test dynamic models. I understand this is a niche thing, so let me give an example of how these might work.

Tools: Python, `Mesa` (ABM modelling library in Python) or just `pyplot`+`pandas` to keep it simpler.

Suppose for simplicity that our $\boldsymbol{\theta}$ turns out to be a vector of two parameters: the length of the user's post (correlated negatively with their popularity) and the emotional sentiment of the user's post (correlated positively with their popularity). We know these things correlate with the user's popularity but we do not know what precisely is the causal mechanism underlying this correlation (assuming there is one). We hypothesize one possible causal mechanism: other users in the network simply have limited time that they can devote to consuming online content and they happen to get the most bang for their buck by consuming content of short and punchy form, thus elevating it to a viral status.

We then formalise the social media space as a directed network, where nodes are users and user i is connected by a directed edge to j iff i follows j. We then formalise the causal mechanism: say, users can only spend so

much time in one online session, so they can only choose to spend some limited time x. For every post they read, they gain some utility u depending on how entertaining it is and expend some energy e depending on how long it is. So, for instance, a long post might be high in both u and e, while a short but sensational post may be moderately high in u but very low in e.
The users' goal is to maximise u and minimise e within x, so they choose who to follow accordingly. Every round, they experiment with who they follow and read until they find the optimal setup. It then may (or may not!) turn out that reading posts that are short and sensationalist is simply the most optimal way for the users to satisfy their natural desire to enjoyably spend their limited time x. Note the interesting bit about this explanation: the causal mechanism is completely innocent! Looking at the parameters of interest, we could have stipulated something like 'oh, social media users are just stupid, so they read short and sensationalist stuff' – but in our model, users are not stupid, they have absolutely natural and justified preferences.

Another benefit of this exploration is that it gives us the idea about the effect of the structural component (i.e., the topology of the model) affects the results. So, for instance, it may be that some causal mechanism may produce the desired effect on social media platforms where users are very well-connected (e.g. Twitter) but not on social media platforms where users are highly clustered (e.g. Telegram). This is also a piece of the causal puzzle.

But this is only if we have time and desire.