Web Science weekly project wk 3-4 – Network dynamics

November 2022.

Teacher/supervisors: Tiago Prince Sales and Claudenir Morais Fonseca
TA/technical support: Boris Belchev, Plamen Bozov, Cosmin Ghiauru, Puru Vaish

Material:
Easley & Kleinberg, Networks, Crowds and Markets, 2010

- Chapter 16, Information Cascades
- Chapter 17, Network Effects
- Chapter 18: Power Laws and Rich-Get-Richer Phenomena
- Chapter 19: Cascading Behavior in Networks

Project prerequisites:

- Python as a programming language
- YouTube Data API from Google
- Spotify data set and You data set

Project 'Music'

In this project we study music preferences of a large audience. We collect the data on music videos (songs) from the *Spotify* and *YouTube* websites, using their API. Our objective is to learn how to collect such data and investigate whether information cascading effects, network effects, rich-get-richer phenomena and cascading behavior phenomena play a role in people's music preferences.

Provided data sets

We provide a *Spotify* data set. This data set consists of snapshots of the *Spotify* top-100 songs over a period of about one year, with one snapshot per day (taken at the same time each day). We also provide a *YouTube* data set. This data set contains information on 100 distinct *YouTube* songs over the same period of about one year, with daily information (taken at the same time each day) per song. The tracked songs in the *YouTube* data set are the same as those which appear in the first snapshot of the top-100 in the *Spotify* data set. In addition, we provide two further *YouTube* data sets. These data sets contain information on songs that have previously been announced by two radio stations as songs with hit potential. One data set contains *Radio 3FM* 'megahit' songs and the other *Radio 538* 'alarmschijf' songs. In both cases the covered period is much shorter, about 2 weeks.

The data sets are available on Blackboard as JSON files named *spotify_top100*, *youtube_top100*, *radio3fm_megahit*, and *radio538_alarmschijf*.

In the assignments below you are free to choose parts of the available data sets (specific songs or covered periods) to illustrate (the absence of) certain effects and phenomena.

Assignments

1. Cascading effects

Read the material in Chapter 16 on cascading effects. Plot the difference between the number of likes and dislikes for several songs.

Do you think we observe cascading effects? Is there a difference between songs that are already popular (in the top-100) and those that are not (megahit or alarmschijf)?. The model in Chapter 16 in its pure form cannot be applied to music preferences. Why not? Can you modify the model accordingly? Can you quantify cascading effects in this setting?

There is no one right answer right to this question, and the answer may depend on a particular song. Make your conclusions based on your data and present clear arguments.

2. Network effects

Study the material in Chapter 17. Our goal now is to investigate whether we observe network effects in music preferences. How will you choose the data for this purpose? Which songs will suit most?

For several songs of your choice plot the actual number of views against time. Assume that without network effects we can expect that users visit a website with a certain frequency and view the song if it matches their taste. Hence, the expected number of views grows linearly in time. Compare your plot to Figure 17.4. Do you think you observe network effects?

How will you interpret the network benefit function f, the intrinsic interest function r, and the price p^*? Try to specify the model that best fits the data.

3. Rich get richer (popularity) effects

Study the material in Chapter 18. We investigate whether rich-get-richer phenomenon explains the dynamics of the number of views.

Plot the distribution of the number of views among the songs on several different days. Do you observe power laws?

Assume that the number of views in the next day is proportional to the total number of views up to the day before. Argue that in this case the number of views will grow exponentially in time.

Look again at the plots for the number of views over time that you produced in assignment 2. Do we observe exponential growth on data? Maybe we observe exponential growth at least some periods of time? Do you think we observe the rich get richer phenomenon?

Compare the ranking of songs in the *Spotify* data set (based on their position in the top-100) with the ranking of songs in the *YouTube* data set (based on their number of views) over time. Are the outcomes in line? Why (not)?

4. **Information diffusion**

Read Chapter 19. Do you think the model of information diffusion applies for music preferences? How can you observe this on the data? Is it related to other phenomena discussed above? Again, there is no one right answer to these questions, try to formulate your own ideas.

5. **Create your own data**

Use the *YouTube* API to collect your own data set to investigate distribution of popularity and the long tail phenomenon. Start with obtaining the number of views of a song of your choice. Make a random selection from the recommendations that come with this song, and obtain the number of views of this recommended song. Repeat this sequence at least 100 times. Does this data set illustrate the strong variation in the market share of different songs (stronger than would be expected on basis of a normal distribution)? Can you observe the long tail in the distribution of popularity? Visualize the distribution with a graph similar to Figure 18.4.

A document with instructions on how to use the *YouTube* API is available on Blackboard. The *YouTube* API will return results in JSON format, so your data set will be a JSON file.

6. **Conclusions**

Make conclusions: 1) which effects and models explain best the given data on music preferences, 2) which data we need ideally if we want to investigate, respectively, cascading, network, rich-get-richer effects, and information diffusion in music preferences?

Deliverable

Write a report which contains the results requested in the assignments. The report has 6 sections corresponding to the 6 assignments. At least the following information should be provided:

- Plots (in sections 1, 2, 3 and 5)
- Explanations of the Python code (in sections 1, 2, 3 and 5)
- Answers to the questions, where you discuss the data and reason about observed or expected effects using the theory presented in the mentioned chapters of the book (in all sections)

Your Python code and JSON data set should be provided separately from the report.

Upload all results in a single Zip file.

Grading

The relative weight for grading of the assignments/sections is as follows:

- Assignments/sections 1, 2, 3, 5: 20%
- Assignments/sections 4, 6: 10%

Only the presence of the above-mentioned information in the report of course is not enough for a maximal mark. The report will be also evaluated for its readability, the suitability and consistency of the solution, the coverage (depth) and the reflection on the results (choices and limitations).

Deadline

Upload a single Zip file containing the report (in PDF file format), the Python code used, and the JSON data set created. The file should be uploaded to Canvas before Friday 9 December 23:59 hrs.