UDK 004.94:519.72 doi: 10.26906/SUNZ.2021.4.048

E. Ivohin [1], V. Gavrilenko[2], P. Vavryk [1]

[1] Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
[2] National Transport University, Kyiv, Ukraine

ONE METHOD FOR ESTIMATION AUDIENCE OVERLAP IN SOCIAL MEDIA

Abstract. In this paper we provided the definition of the Audience overlap network, as well as proposed a simple algorithm to compute overlap between two users on social media based on public data about their followers. There was proposed an alternative approach for computing overlaps based only on public data about users. This approach allows to include content overlap and activity patterns signals to be incorporated into more general statistical models featuring other covariates such as influencers' direct engagement in shared conversations; relative influencer sizes and histories and links to similar third-party content to recover otherwise censored network structures and properties. For validate results there was designed a calibration process which utilizes Evolution Strategies algorithm to find a set of conditions which will make Audience overlap network built using similarity measures structurally equivalent to the Audience overlap network build on full information about followers.

Keywords: content, overlap, social media, algorithm.

Introduction

With growing number of users in recent years, social media platforms have become not only the prime place for public discourse, but also a main source of news for many people. This massive development increased a need of their analysis on structural and content levels. Every level has its own state of the art tools and methods of research. But in many cases, understating of influence on social media requires performing analysis on both levels simultaneously and mapping "influencers" and ordinary users serving as their audiences. To map social media influencers to their audiences we explore a general approach for building Audience overlap networks (AONs). AONs can be used for multitude of applied problems e.g., detecting groups of users spreading disinformation, coordinated messaging campaigns, general community detection, etc. AONs can be built for most of modern social media platforms (e.g., Facebook, Twitter, or YouTube), the only condition we require for these platforms is to have a follower-followee connections structure. While platforms like Twitter or Reddit grant unrestricted access to lists of followers or members of influential channels and profiles, others like YouTube or Facebook do not grant similar access or otherwise restrict its granularity. This puts additional pressure for general approach of building AONs but at the same time opens opportunity for exploration of deeper connections between influencers and their audiences.

In this paper, our aim is to provide a general definition of the AON, data processing steps needed for creating AON and the challenges faced when building AONs for social media platforms with minimal amount of public data available.

To demonstrate challenges faced while working with limited input data we first use Twitter to provide a baseline AONs which can be used for any out-of-sample testing.

Our first pick was Twitter because not only Twitter holds the position of one of the most popular social networks (most official government accounts are present there e.g., President Biden, United Nations, World Health Organization) but also Twitter was designed in way that all interaction between users remain visible to everyone else. Twitter also has a flexible public application programming interface (API).

We will follow definitions in [1], where audience overlaps were computed by link and cross-link similarity.

Data processing

In this section we describe the topics and the process of collecting data for building AONs (Table 1).

Table 1–**Topics and influencers**

ID	Name	Influencers
T_1	Democratic primaries	Joe Biden, Mike Bloomberg, Pete Buttigieg, Kamala Harris, Amy Klobuchar, Bernie Sanders, Elizabeth Warren
T_2	Airline geeks	Boarding Area, French Painter, One Mile at a Time, Runway Girl, Secret Flying, Wander Me
T_3	Global issues	Inside Climate, Ro Khanna, Sierra Club, The Economist, World Bank, Yale E360, Chris J. Zullo
T_4	Technology investors	Adam Scrabble, Epsilon Theory, Eric R. Weinstein, Fast Company, Naval, Nick Timiraos, WIRED

Topics and influencers

We consider four topics and influencer groups:

T_1: Democratic primaries. Conversations about the 2020 Democratic Party presidential primaries. Influencers are selected from the top presidential contenders.

T_2: Airline geeks. Conversation related to the airline industry and business air travel. Influencers include reviewers of frequent flyer programs, travel experience bloggers and industry insiders.

T_3: Global issues. Conversations about reporting, analysis and opinions on global environmental, climate and energy issues. Influencers include magazines like The Economist, and selected journalists and organizations like The World Economic Forum.

T_4: Technology investment. Conversations on technology investment, emerging technologies and entrepreneurship. Influencers include prominent venture capitalists and industry publications.

Influencers were selected for both relevance and focus on a topic and ability to generate engagements with their content. Table 1 lists selected influencers for every topic in alphabetical order for entities and last name for individuals.

In most of the cases the analysis of social networks is more effective when applied to specific subset of conversation (posts) and authors (users).We define social media topic as all authors and conversations which mention specific keywords. In our case keyword can represent a name or multiple spellings of author as well as some general expressions related to the area of interest.

For every topic, we used the Twitter API [2] to query and download all tweets for the first half of 2020, including tweets by identified influencers; then download the list of followers for every influencer.

Audience overlaps network

We define AON as graph with influencers as vertices, where the weight of the edge between vertices (influencers) defined as the number of shared followers between them (audience overlap).

Consider Fig. 1 as a visualization of the structure of some topic X. Here we have identified influencers A and B and eight followers. We calculate the audience overlap between A and B as

$$O(A, B) = \frac{|F_A \cap F_B|}{|F_A|},$$

where F_A and F_B denote sets of followers of A and B respectively. In this particular case $O(A, B)$ = 3/6 and $O(B, A)$ = 3/5. After downloading all followers for each influencer, we built the AON for every topic see Fig. 2.

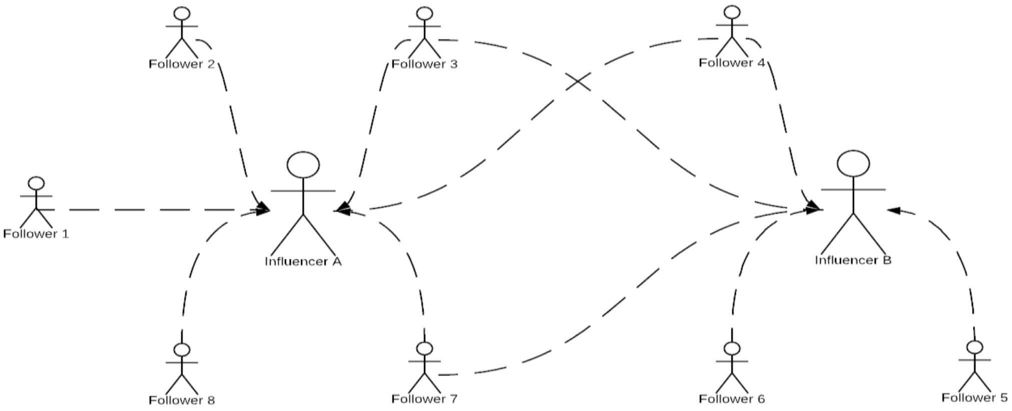
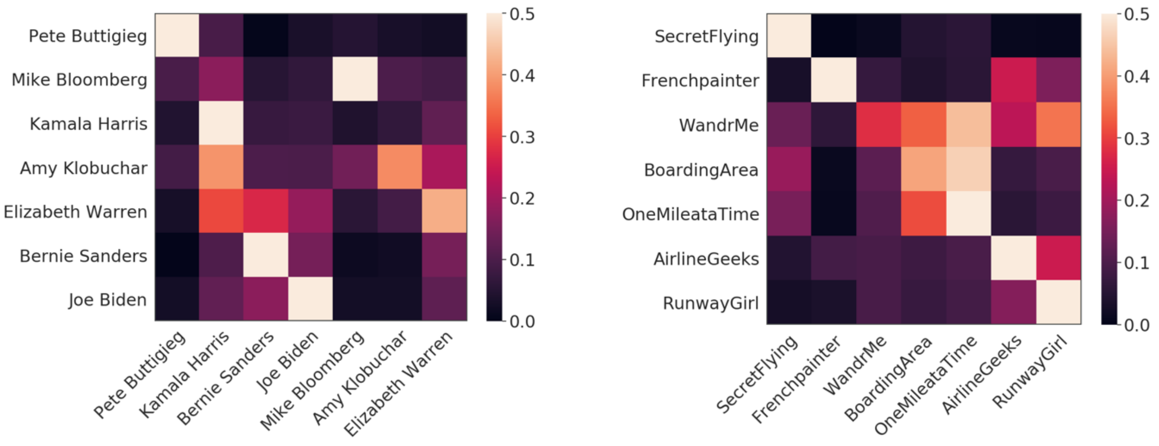
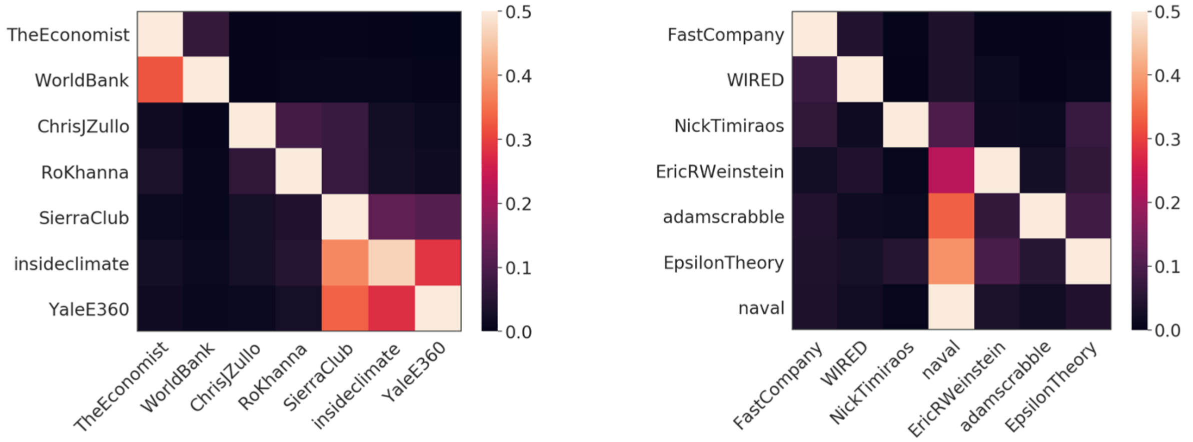


Fig. 1. A sample follower-followee network for influencers *A* and *B*



(a) AON for T_1

(b) AON for T_2



(c) AON for T_3

(d) AON for T_4

Fig. 2. Visual presentation of AON for every topic in a table form where every cell is an average percentage of shared followers between two influencers.

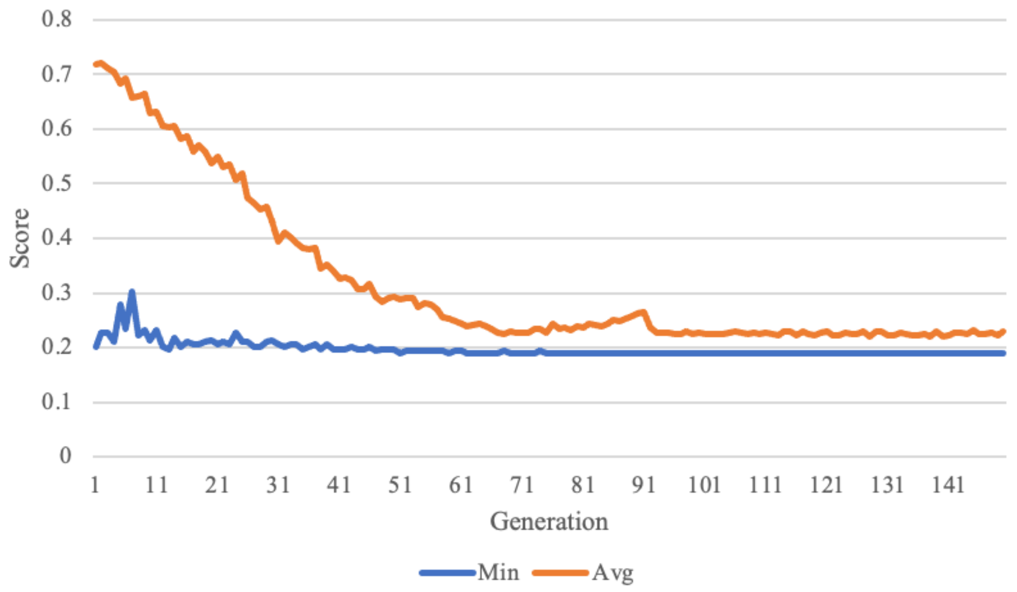


Fig. 3. Calibration progress for the topic T_1.

Lack of data

The process of building AONs based on Twitter data raises no additional problems, at same time many large social media platforms have restricted public access to information about follower. This is the case with Facebook, even though the follower-followee structure is present on the platform we are unable to build AONs. At the same time Facebook still leaves a lot of information about influencers and their activity including content of posts, reactions, comments, all annotated with timestamps.

We defined a set of functions, each of them comparing similarity of two influencers based on some characteristic:

- $S_1(A, B)$ – compares similarity between content of A and B using cosine distance based on the features produced by Bag-of-Words method [3].
- $S_2(A, B)$ – performs a search for links to same resources ("http://*", "https://*"", "www.*", etc.) published by both A and B. This function also checks for domains similarity.
- $S_3(A, B)$ – compares distributions of publishing activities of A and B.
- $S_4(A, B)$ – compares distributions of comments under publications of A and B.

We consider that every comparison function has different weight within different topics, so the final comparison function is defined as

$$S(A,B)=\sum_{i=1}^{4} w_i * \max(0,(S_i(A,B)-b_i)),$$

where $S_i(A, B)$, $i=1,2,3,4$ - values defined above.

The solution for building AONs on partial data is to compare influencers using function $S(A, B)$ instead of comparing actual followers. But having multiple unknow parameters in comparison function doesn't instantly provide AON, but rather requires understating of the communities' structure from the platform with known followers.

Calibration

After we built AON for every topic of interest, we tried to find the set of values for weight parameters so that the AON built using *S(A,B)* function will structurally match baseline Twitter AON.

For this purpose, we designed a rigorous calibration process which utilizes Evolution Strategies algorithm [4] and performs a search over values of weight parameters to minimize structural difference between output and baseline AON.

We selected the following measurements of structural difference:

- Number of clusters (> 2, > 3, > 4, > 5).
- Diameter of the graph.
- Betweenness centrality.

To get more clear calibration results, we extended the list of influencers in every topic to 30.

Results

Figure 3. presents calibration results for the topic T_1 after 150 generations of the algorithm.

The optimization started optimistically by improving average score by factor of 3 in just 50 generations, but after 100 generations the algorithm couldn’t find a better set of parameters.

The same problem appeared in calibrations for the rest of the topics.

This is a clear indication that the function for measuring similarity can find only macro-level differences leaving micro-level differences unexplored.

This problem can be solved by applying state of the art Natural Language Processing techniques for measuring content similarity [5] or by using methods for analyzing information flows in social networks (Social Network Mining) [6].

Summary

We identified the role of Audience overlap networks as a powerful tool for analysis of social media.

We described a general approach for building AONs while having full understating of follower-followee relations.

Using four topics, we showed that it is possible to build an approximation of AON using only minimal amount of input data.

We are also certain that by applying better content similarity measures it is possible to improve process of building AONs and remove the need in knowing full structure of the followers.

Список літератури

1. Subhayan Mukerjee, Sílvia Majó-Vázquez, and Sandra González-Bailón. Networks of Audience Overlap in the Consumption of Digital News// Journal of Communication.- 2008. - V.68. – Iss.1. – Pp.26–50.
2. Twitter API v1.1. [Online] – Available from: https://developer.twitter.com/en/docs/twitter-api/v1.
3. Zhang Y, Jin R, Zhou Z H. Understanding bag-of-words model: A statistical framework. // International Journal of Machine Learning and Cybernetics. – 2010. - 1(1). – Pp.43-52.
4. Beyer, Hans-Georg, Schwefel, Hans-Paul. Evolution strategies - A comprehensive introduction// Natural Computing. – 2002. - 1(1). – Pp.3-52.
5. Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, et al. Universal sentence encoder// arXiv preprint, 2018. arXiv: 1803.11175.
6. Гусарова Н.Ф. Анализ социальных сетей. Основные понятия и метрики. – СПб.: Университет ИТМО, 2016. – 67 с.

Received (Надійшла) 10.11.2021
Accepted for publication (Прийнята до друку) 24.11.2021

Про один метод оцінки перетину аудиторій у соціальних мережах

Є. В. Івохін, В. В. Гавриленко, П. Р. Ваврик

Анотація. У цій статті пропонується загальне формулювання мережі перетину аудиторій в соціальних мереж, а також простий алгоритм для визначення перетину аудиторії двох окремих користувачів, який базується на використанні публічних даних про їх послідовників. Запропоновано альтернативний підхід пошуку перетину аудиторій, який базується на схожості користувачів на основі лише загальнодоступних даних. Підхід дозволяє розглядати сигнали схожості контенту і особливостей поведінки для застосування у загальних статистичних моделях, що формалізують типові для мереж характеристики (коваріати), такі як пряма участь впливових осіб в загальних обговореннях; відносні розміри і історії впливових осіб, а також посилання на аналогічний сторонній контент для відновлення цензурованих мережевих структур і властивостей. Для валідації результатів застосовано процес калібрації і встановлено наявність залежності перетину аудиторій користувачів від схожості створеного ними контенту та особливостей їх поведінки.

Ключові слова: контент, перекриття, соціальні мережі, алгоритм.