Dynamic Topic Modeling using Social Network Analytics

Shazia Tabassum[1[0000−0003−0782−7054]], João Gama[1[1111−2222−3333−4444]], Paulo Azevedo[1[2222−−3333−4444−5555]], Luis Teixeira[2[2222−−3333−4444−5555]], Carlos Martins[2[2222−−3333−4444−5555]], and Andre Martins[2[2222−−3333−4444−5555]]

[1] INESC TEC, University of Porto, Rua Dr. Roberto Frias, Porto, Portugal
https://www.inesctec.pt/
[2] Skorr, Portugal
https://skorr.social/

Abstract. Topic modeling or inference has been one of the well-known problems in the area of text mining. It deals with the automatic classification of words or documents into similarity groups also known as topics. In most of the social media platforms such as Twitter, Instagram, and Facebook, hashtags are used to define the content of posts. Therefore modelling of hashtags helps in categorising posts as well as analysing user preferences. In this work, we tried to address this problem involving hashtags that stream in real-time. Our approach encompasses graph of hashtags, dynamic sampling and modularity based community detection over the data from a popular social media engagement application. The approach seems to dynamically produce sensible clusters of topics.

Keywords: Topic modelling · Social network analysis · Hashtag networks.

1 Introduction

Social media applications such as Twitter, Facebook, Instagram, Google, linkedin have now become the core aspect of people's lifes. Consequently, these are growing into a dominant platform for businesses, politics, education, marketing, news and so forth. The users are interested in which of such topics or products is one of the primary questions of research in this area. Inferring topics from unstructured data has been a challenging task.

Typically, the data gathered by the above applications is in the form of posts generated by the users. Posts can be short texts, images, videos or messy data. Classification of posts into topics is a cumbersome problem. While topic modeling algorithms such as Latent semantic analysis (LSA) and Latent Dirichlet Allocation (LDA) are originally designed to derive topics from large documents such as articles, and books. They are often less efficient when applied to short text content like posts [1]. Posts on the other hand are associated with rich user-generated hashtags to identify their content, to appear in search results and to enhance connectivity to the same topic. In this work, we propose to use

these hashtags to derive topics using social network analysis methods, mainly community detection.

Moreover, the data generated from social media is typically massive and high velocity. Text mining based models are difficult to cope up with the scale and velocity of data in its entirety. Therefore we tried to address the above issues by proposing an approach with the contributions stated below:

1. We propose fast and incremental method using social network analytics
2. Unlike conventional models we use hashtags to model topics which saves the learning time, preprocessing steps, removal of stop words etc.
3. Our model categorises tags/words based on connectivity and modularity. In this way the tags/words are grouped accurately even though they belong to different languages or new hashtags appear.
4. We employ dynamic sampling mechanisms to decrease space complexity

2 Related Work

Word2Vec [5] is one of the most popular word representation techniques. This model outputs a vector for each word, so it is necessary to combine those vectors to retrieve only one representation per product title or post, since there is the need to have the entire sentence representation and not only the values of each word. Word2Vec output dimensions can be configurable, and there is no ideal number for it since it can depend on the application and the tasks being performed. Since this type of model is very common and can be expensive to train, we will not use in our experimental section...

3 Case Study

An anonymized dataset is collected from a popular social media activity management enterprise. The dataset ranges from January to May 2020; comprises of 1002440 posts with 124615 hashtags posted by users on different social networking platforms (Twitter, Facebook, Instagram, Google). The content of posts is not available, instead the posts are identified with posts IDs and the users are identified with anonymous user IDs. Figure 2 displays the distribution of topics over users. A few topics are discussed by large number of users and more topics are discussed by only a set of few users. This satisfies a power law relation which is usually seen in most of the real world social networks [7]. Each post can include one or more hashtags or none. The number of posts per day is given in Figure 1. As one can observe there is decreased activity on weekends (Saturday and Sunday) compared to other days; with the peaks on Fridays. The data in the last week of April had not been available which can be seen as an inconsistency in the curve with abnormally low activity close to zero. Figure 3 displays the top ten trending hashtags in the given dataset. This type of analysis with the help of topic modelling or trending hashtags can be used to detect events. In the figure, five of the top frequent hashtags are relating to Covid19. What we need from

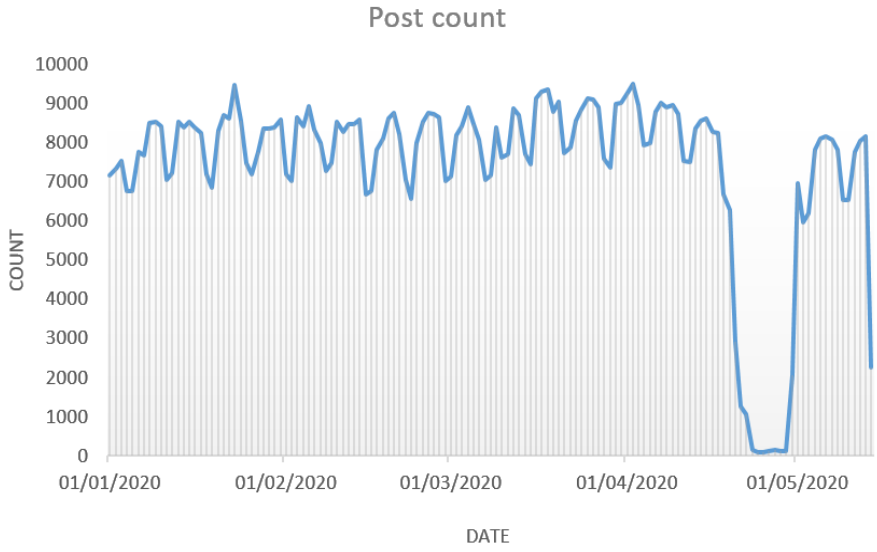

Fig. 1. Users post activity per day

our model is to cluster all these hashtags and the one's that are less frequent to be classified as one topic relating to Covid19. Similarly with the other tags and their related posts. In order to achieve this we followed the methodology briefed below.

4 Methodology

Text documents share common or similar words between them, which is exploited in calculating similarity scores. However, topic modeling in hashtags is unlike documents. Therefore, here we considered the hashtags to be similar based on their co-occurence in a post.

The first step in the process is to build a co-occurrence network from the streaming hashtags incrementally. The hashtags that needs to stay in the network are decided based on the choice of the sampling algorithm in Section 4.3. There after the communities are detected in the network as detailed in Section 4.4.

4.1 Problem Description

Given a stream of posts $\{p_1, p_2, p_3...\}$ associated with hashtags $\{h_1, h_2, h_3...\}$ arriving in the order of time, our approach aims to categorize similar posts or hashtags into groups or clusters called topics at any time t. Each post can be associated with one or more hashtags.

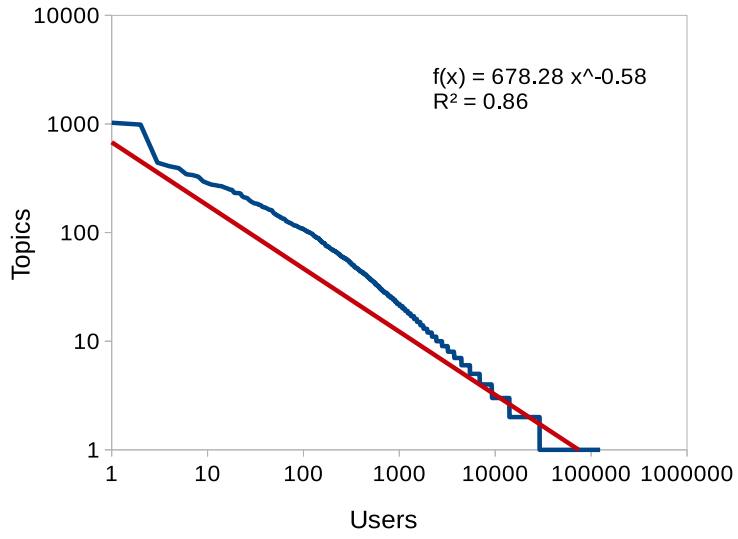

Fig. 2. Users vs topics distribution (blue line). Power curve following given function(red)

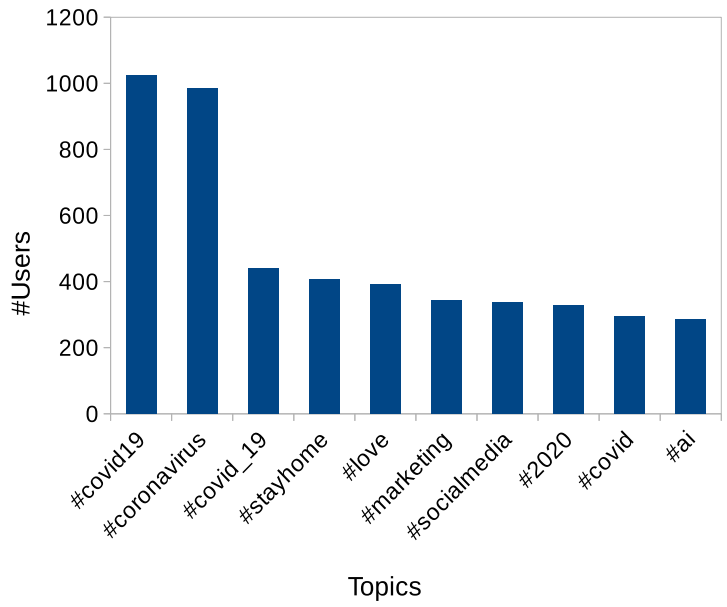

Fig. 3. Top ten trending hashtags over time

4.2 Hashtag Co-occurence Network

In our graph based approach, we constructed the network of hashtags by creating an edge $e = (h_i, h_j, t)$ where $i, j \in \mathbb{N}$ between the ones that has been tagged together in a post .

4.3 Stream Sampling

As posts are temporal in nature, generating in every time instance, so are the hashtags. Also, there are new hashtags emerging over time. Moreover, the context for grouping hashtags may change over time. For example, hand sanitisers and face masks were not as closely related as with the onset of covid19. Therefore, we employed the approach of exploiting the relation between hashtags based on the recent events or popular events by using the realtime dynamic sampling techniques below.

Sliding Windows Sometimes applications need recent information and its value diminishes by time. In that case sliding windows continuously maintain a window size of recent information. It is a common approach in data streams where an item at index i enters the window while another item at index $i - w$ exits it [3]. Where w is the window size which can be fixed or adaptive. The window size can be based on number of observations or time. In the later case an edge (h_i, h_j, t) enters window while an edge $(h_i, h_j, t - w)$ exits.

Space Saving The Space Saving Algorithm [4] is the most approximate and efficient algorithm for finding the top frequent elements from a data stream. The algorithm maintains the partial interest of information as it monitors only a subset of items from the stream. It maintains counters for every item in the sample and increments its count when the item re-occurs in the stream. If a new item is encountered in the stream, it is replaced with an item with the least counter value and its count is incremented.

Biased Random Sampling This algorithm [6] ensures every item m goes into the reservoir with probability 1. An item n from the reservoir is chosen for replacement at random. Therefore, on every item insertion, the probability of removal for the items in the reservoir is $1/k$, where k is the size of reservoir. Hence, the item insertion is deterministic but deletion is probabilistic. The probability of n staying in the reservoir when m arrives is given by $(1 - 1/k)^{(m-n)}$. As m increases, the probability of n staying in reservoir decreases. Thus the item staying for a long time in the reservoir has an exponentially greater probability of getting out than an item inserted recently. Consequently, the items in the reservoir are super linearly biased to the latest time. This is a notable property of this algorithm as it does not have to store the ordering or indexing information as in sliding windows. It is a simple algorithm with $O(1)$ computational complexity.

4.4 Community Detection

Community detection is very well known problem in social networks. To understand the community detection problem the reader should know what is a community in terms of a social network. Communities are groups, modules or clusters of nodes which are densely connected between themselves and sparsely

connected to the rest of the network. The connections can be directed, undirected, weighted etc. Similarly, the context of a relation can also be different. The communities can characterize a group of nodes as friends when the link type is "friend of". Similarly, in metabolic networks the proteins can be grouped based on the similarity of functions, where the link type is "similar function". Therefore, the type of connection defines the context of the community formation. A network of same nodes can form different communities when the relation type is different. Nevertheless, the links can be straightforward like friendship on facebook or it can be derived, example similarity as stated above. The communities can be overlapping (where a node belongs to more than one community) or distinct. Community detection is in its essence a clustering problem. Thus, detecting communities reduces to a problem of clustering data points. It has a wide scope of applicability in real-world networks.

In this work, we applied the community detection algorithm proposed by [2] on every dynamic sample snapshot discretely. However, an incremental community detection algorithm can also be applied on every incoming edge. Nevertheless, the method used in [2] is a heuristic based on modularity optimization with a fast runtime of $O(n.log_2 n)$, where n is the number of nodes in the network. In our case n is very small compared to the total number of nodes in the network, for instance n is equal to the number of hashtags in a sliding window. A resolution parameter controls a high or low number of communities to be detected.

5 Experimental Evaluation

The experiments are conducted to evaluate the above method of detecting topic clusters in a real world data from social media applications as detailed in Section 3. To facilitate visual evaluation and demonstration, the size of samples is fixed to be 1000 edges. The resolution parameter in community detection for all the methods is set to default i.e 1.0. The detected clusters are shown in Figures 4, 6, and 5. The figures represent sample snapshots in the end of stream. Sliding windows and biased sampling considers repetitive edges as the frequency or weight of an edge which is depicted as thick arrows or lines in the figures. The thicker edge represent stronger connection between two hashtags.

We see that the clusters in the figures clearly make sense. Each cluster with a different color in the figure represents a topic. However, the clusters formed by sliding windows are more denser than the other two. Quantitative metrics of these graphs are displayed in Table 1. The bias to low degree hashtags has increased the number of components and decreased the density. Nevertheless, a large cluster of the popular topic covid19 can only be seen in space saving because sliding window and biased sampling collect data from the end of stream that is from the month of May, where it has low occurrence. The posts and users relating to these hashtags can be further investigated for numerous applications.

The choice of sampling algorithm has different trade offs. For finding the most frequent or trending topics from the stream over time, Space Saving is a relavant choice; however, it is computationally expensive compared to the

other two though it is space efficient and the fastest one of its genre. The least time complexity one among the three is Biased sampling but lacks in terms of structure in this case, with a very sparse graph.

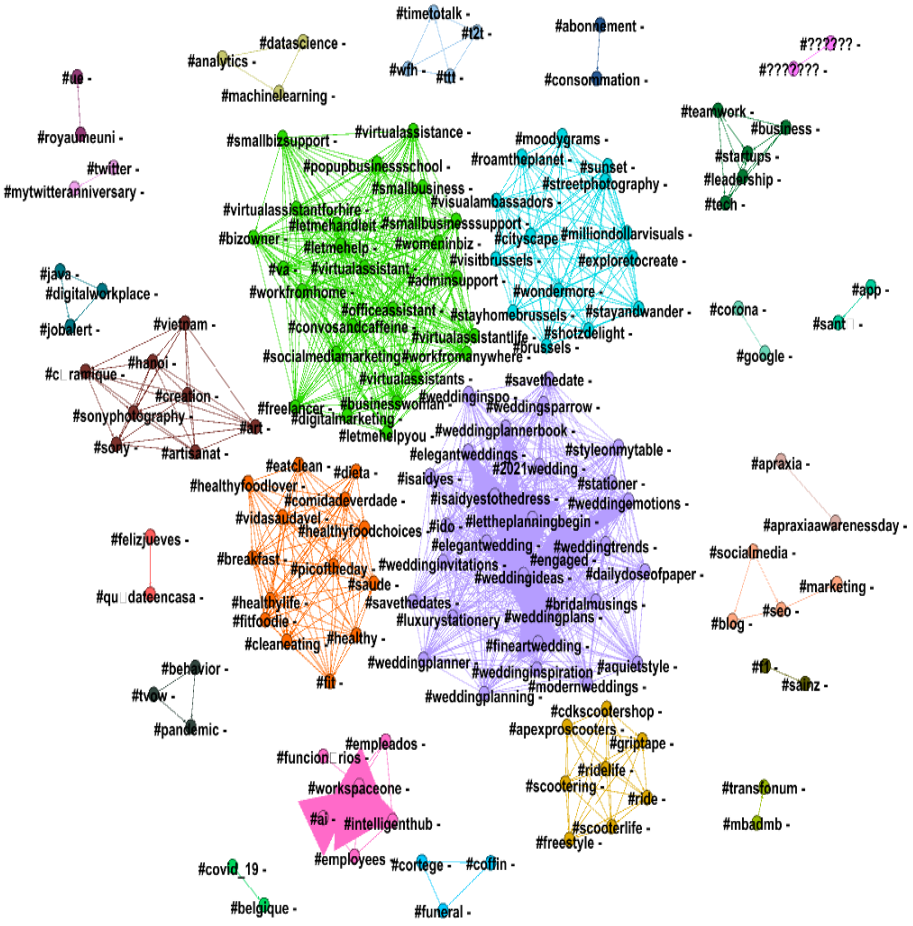

Fig. 4. Sliding Window

6 Conclusion and Future Work

In this work we have presented a fast and memory efficient approach of incrementally categorising posts into topics using hashtags. We proved the efficacy of method over a real world data set comprising of different social media applications data. We discussed on how the different sampling algorithms can effect

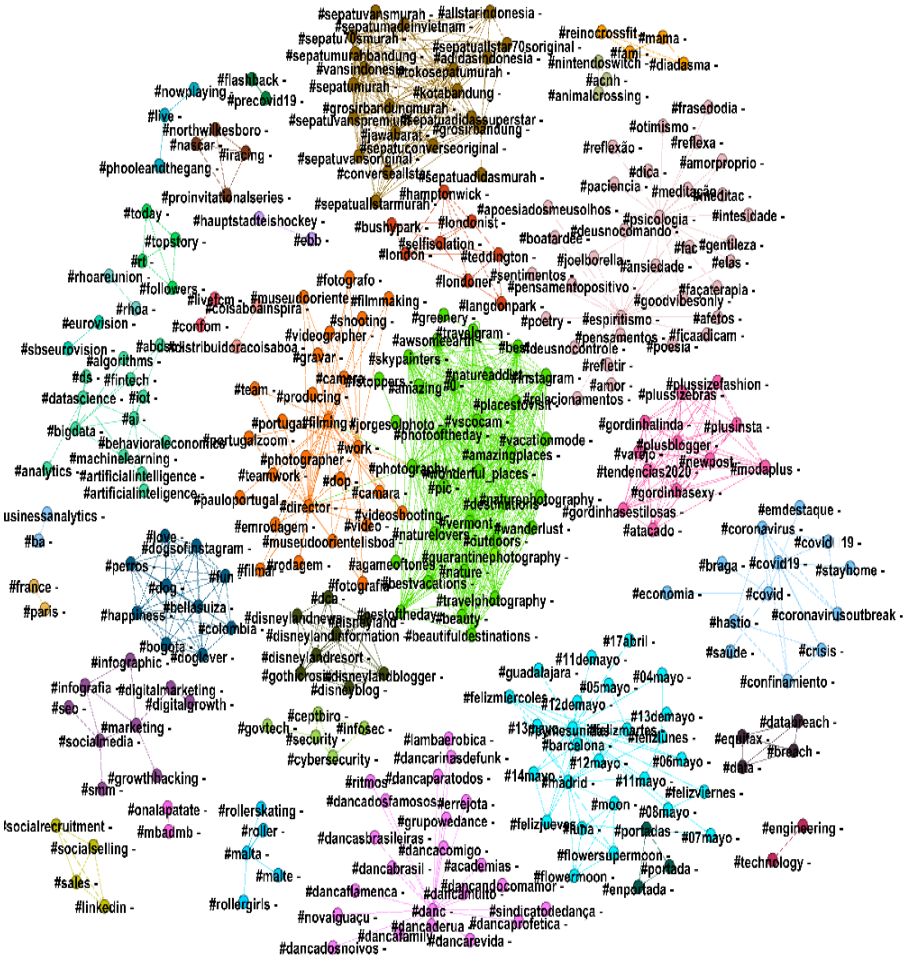

Fig. 5. Space Saving

	Average degree	Avg. weighted degree	Density	Modularity	#Clusters
Sliding window	6.6	6.667	0.087	0.723	25
Space saving	3.413	3.413	0.022	0.806	33
Biased sampling	2.687	2.747	0.015	0.872	61

Table 1. Network properties

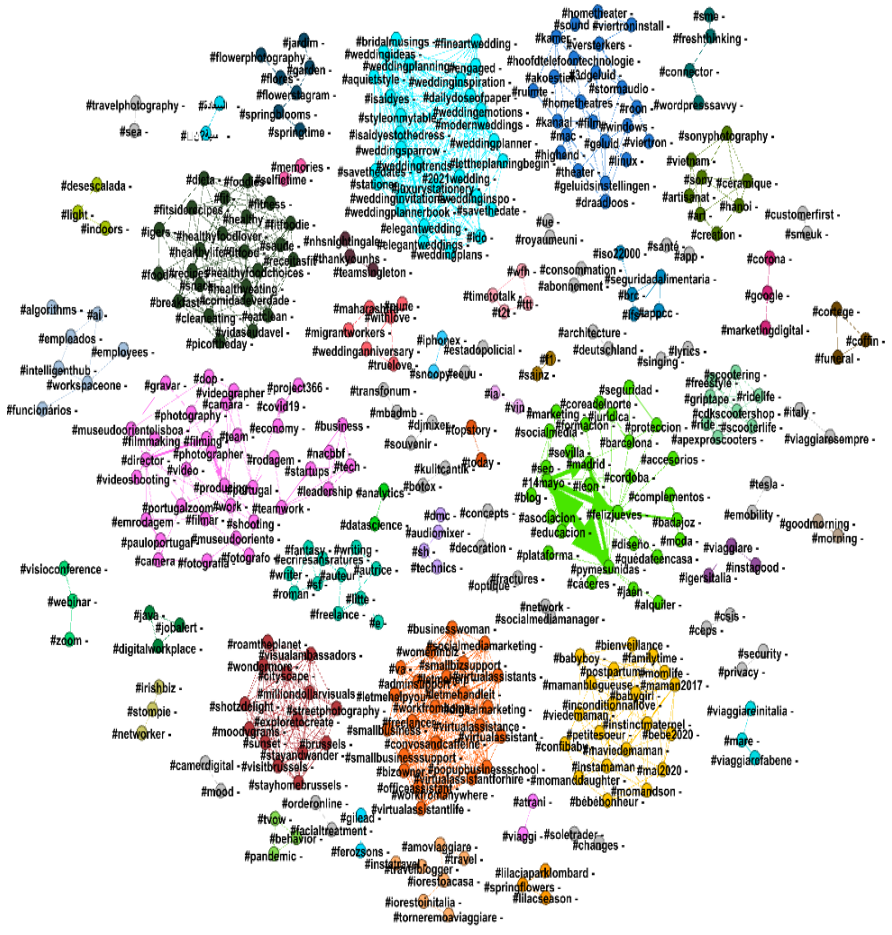

Fig. 6. Biased Random Sampling

the outcome. Further, we considered their biases and trade offs. We analysed the seasonality and trending tags in the data. We compared their outcome in terms of quality and structure of clusters. To facilitate comprehensibility we preferred network visualisation layouts to present the results over the conventional presentation using tables.

There can be many potential applications as an advancement of this work. The users posting in particular topics can be classified accordingly to analyse their preferences for product marketing and identify nano influencers to enhance their engagement. On the availability of posts text we can implement other topic models and improve them using our approach. Additionally, we intend to analyse the trend of topics overtime and the evolution of communities. Besides, predicting hashtags for the missing ones using our topic model.

Acknowledgements

References

1. Alash, H.M., Al-Sultany, G.A.: Improve topic modeling algorithms based on twitter hashtags. In: Journal of Physics: Conference Series. vol. 1660, p. 012100. IOP Publishing (2020)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
3. Gama, J.: Knowledge Discovery from Data Streams. Chapman and Hall / CRC Data Mining and Knowledge Discovery Series, CRC Press (2010)
4. Metwally, A., Agrawal, D., El Abbadi, A.: Efficient computation of frequent and top-k elements in data streams. In: International conference on database theory. pp. 398–412. Springer (2005)
5. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. pp. 746–751 (2013)
6. Tabassum, S., Gama, J.: Sampling massive streaming call graphs. In: ACM Symposium on Advanced Computing. pp. 923–928 (2016)
7. Tabassum, S., Pereira, F.S., Fernandes, S., Gama, J.: Social network analysis: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(5), e1256 (2018)