Representing Emoji Usage using Directed Networks: A Twitter Case Study

Halley Fede, Isaiah Herrera, S.M. Mahdi Seyednezhad, Ronaldo Menezes

Abstract In online social media, people use emojis to reduce the ambiguity of short texts and to express their feelings in a more clear way. Some text messages contain more than one emoji, and this brings the idea that the sequence of emojis may have useful information that can help us better understand user behavior. One method to analyze the sequence of emojis is to study a directed network of emojis that emerges from the actual sequence for many users. In this paper, in addition to extract a simple undirected co-occurrence network and analyze its corresponding main statistical properties, we build and analyze a directed co-occurrence network from various datasets collected from Twitter. The results show that the distributions in directed network are not random and follow a truncated power-law distribution. Furthermore, the important emojis for each dataset are conceptually related to the subject of the dataset. Via community analysis, we show that most of the emojis tend to be grouped in the top 4 largest communities. Last, the category-based entropy analysis of communities suggests that regardless of theme, the entropy is somewhat constant across different thematic datasets. This proposes that emojis are not used together just because they are from the same category.

Halley Fede
Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York, USA.
e-mail: fedeh@rpi.edu

Isaiah Herrera
Department of Mathematics, Westminster College, Salt Lake City, Utah, USA.
e-mail: irh0612@westminstercollege.edu

Mahdi Seyednezhad
School of Computing, Florida Institute of Technology, Melbourne, Florida, USA.
e-mail: sseyednezhad2013@my.fit.edu

Ronaldo Menezes
School of Computing, Florida Institute of Technology, Melbourne, Florida, USA.
e-mail: rmenezes@cs.fit.edu

1 Introduction

Emojis are ubiquitous in online communication. They serve many purposes, among which resolving ambiguity in written communication is the most common one. However, their wide adoption also allow people to better express their opinion or feelings on different subjects in a very short and concise way. Social media is a very popular platform to express feelings and opinions online which make it a prime source for emoji use.

Emojis are the small pictographs that have been widely adopted worldwide. They are basically small pictures which users can use to supplement their text and express more clear feelings. Emojis were first deployed in Japan in 1990's, then the Unicode organization standardized [4] it by assigning certain codes to these pictographs; this enabled people to see the same pictographs across different platforms. Table 1 shows the categories (a.k.a. orders) and subcategories of emojis defined by the Unicode organization.

Table 1 Major orders of emojis with samples and sub-orders

Major orders	Samples	Some sub-orders
Smiley & People	😊😂😀	face-positive, face-neutral, face-negative
Animals & Nature	🐵🦊🌵	animal-mammals, animal-birds, plant-other
Food & Drink	🍇🍷🍟	food-fruit, drink, food-vegetable
Travel & Places	🌍✈🚂	place-map, transportation-ground, transportation-air
Activities	🎠🚲⚽	event, sport, game
Objects	📽☎📗	sound, phone, money
Symbols	⛔🚾🔄	transport-sign, arrow, warning
Flags	🇺🇸🇧🇷🇮🇷	flag, country-flag

As previously mentioned, emoji usage has wide adoption worldwide. In 2015, emojis were adopted by Android in addition to iOS; as a result, almost half of the text posted on Instagram contained emojis and it reached to more than 55% of the posts in December 2016 [6]. Evans [5] discusses that emojis can be considered as a system of symbols that have some similarities to languages. All put together, analyzing emojis can give us important information about people using online social media. Since emojis are collected from a large number of short messages as social media posts, its understanding falls under the label of "big data". A common practice to make sense of big data is to apply Network Science approaches to unveil the connection among pieces of the data, following by understanding the communities they form, the distribution of their connectivity, and identification of central pieces [1].

In this paper, we start with an overview of related works in Section 2 followed by a discussion on the datasets we used in this work in Section 3 coupled with how we extract emoji networks from these datasets. Section 4 has two major parts. In the

first one, we visualize the network of emojis extracted from our datasets with the aim of providing visual aid for the understanding of emoji usage. In the second part, we do several network analyses including standard degree distribution analysis and community organization. We finish this section with experiments that look at the organization of these communities using entropy. Finally, we conclude our work in Section 5 and discuss about possible future work.

2 Related work

Before emojis become popular, people already used a sequence of characters named *emoticons* (e.g. "`;-)`", "`;p`") widely. Pavalanathan et al. [11] show that the increasing popularity of emojis is coupled with the demise of their predecessors, i.e. emoticons. More recently, researchers become interested in understanding the meaning and the sequence of the emojis. For example, Barbieri et al. [2] did an extensive study on the meaning of emojis using Twitter data. They worked by looking at their relatedness (the likelihood of two emojis to appear in the same tweet) and their similarities (how much humans consider the emojis to convey the same meaning).

Wijeratne et al.[14] create and analyze a dictionary based on emojis in order to make a machine readable sense inventory for emojis. They use the Unicode, description, image and keywords attached to the meaning of the emoji to create octuples representing the meaning of the emoji. They use open access resources to create their dictionary such as emojipedia and other resources. Finally, they offer a product named "EmojiNet" that is a network of emojis connected based on their meaning and people can search for emojis based on either name or sentiment of emojis.[1]

In the context of networks, Lu et al. [8] analyzed emojis using a network created by point-wise mutual information (PMI). They studied ubiquitous usage of emojis to compare user behavior in different cultures and regions throughout the world. One of their significant findings is that countries with similar emoji usage have similar languages. They only took data from Kika keyboards, while 74.3% of users are under the age of 25.

Seyednezhad et al. [13] created a network of emojis based on their co-occurrence in the same tweet for two different datasets. They claim that the emoji with the maximum edge betweenness can give us a hint about the subject that the tweets were collected. Furthermore, they assume that the degree, edge-weight, and weight-degree distribution can be considered as a structure underlying the emoji usage. Then, they show that those distributions for both datasets are similar and conclude that emojis may have a constant similar structure that may be independent of the subject of messages. They use two datasets of tweets and we want to generalize their work by performing the network analysis on 5 other datasets. Furthermore, we introduce direction and look at how emojis form communities as well as the category organization within these communities.

[1] Search available at http://emojinet.knoesis.org/home.php

3 Data and network

The data for this work comes from tweets collected for different subjects at different time periods. The reason to use this type of data is that we can cover a wider range of data and be more independent of the subject or time the tweets were written. The hope is that such wide set of data minimizes biases from other works. Table 2 shows more details about the datasets.

Table 2 Various datasets collected from Twitter and used in this paper. The subject-area of the datasets covers several areas of interest.

Dataset	Description	# tweets in millions	% Containing emojis	Collection period
G-20	The leaders of the G-20 up to their second level of followers.	10.6	7%	Aug. 24 - Sep. 24, 2014
Organ	Tweets containing organ transplantation terms.	2.5	9%	Oct. 2015 - Apr. 2017
rioSports	A collection of tweets related to sports practiced during the Rio Olympics in Brazil.	1.8	1%	Aug. 05 Aug. 21, 2016
rioTerms	Tweets containing the term Olympics in different languages, also collected during Rio Olympics in Brazil.	5.8	1%	Aug. 05 Aug. 21, 2016
WWC	A collection of tweets during the Women's World Cup 2015 and South American Cup	10.7	1%	Jun. 06 - Jul. 05, 2015
randSample	A collection of about 2 months of random tweets	168.5	less than 1%	Dec. 13, 2016 - Jan. 31, 2017

It is worth noting that we have data related to politics, sports, health as well as a random collection of tweets. The *Organ* collection has the most percentage of tweets containing emojis while the random sample has the least amount.

The main focus of this paper is on creating a *directed network* for each dataset and analyze it. As opposed to previous works [13], we assume the order emojis appearing in a tweet is fundamental and hence better represented using directed edges. For example, if a typical user named Diego writes "I love to eat cake with Marcos ❤️🎂", he wants to say that the act of *eating a cake* is *lovely* when you have a company like Marcos. On the other hand, if Diego tweets "I hate eating cakes 🎂even with my friend Marcos ❤️", it means that he likes Marcos, but apparently he ended up in a situation where he is "forced" to eat a cake. Note that the order of the emojis is related to the sentiment being expressed. In order to build the directed network, we connect each emoji to the emojis following it in the same tweet. Figure 1 shows the process of making directed links between emojis.

4 Experimental Results

4.1 Visualizing Emoji Usage

Before we provide the analyses, we would like to see how the EmojiNets is formed; such visualizations can assist in the understanding of the analyses performed later. It should be noted that emojis are linked if they appear next to each other in a

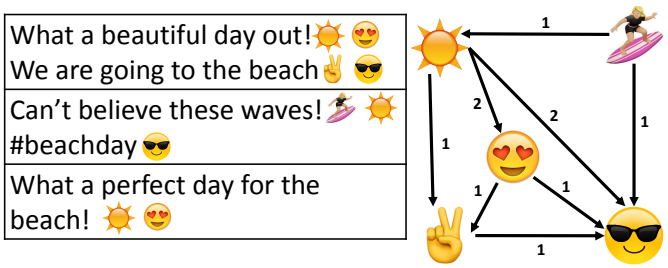

Fig. 1 We create a directed network of emojis by making a connection from emoji to emoji in the order they appear in a tweet. This process is repeated for every tweet in the dataset.

tweet. Such simple and common approach leads to most emojis being connected to each other. Then the network itself will be significantly large and dense. Hence in Figure2 we show just a section of the entire network for the random tweets that is built using a weight-based filter for edges—low weight edges are removed. For illustrative purposes, we can pick two nodes to explain more. For example, a link from 💀 (skull) to 🤣 (rolling on the floor laughing) implies that there are a high number of tweets in our random sample dataset in which users use 💀 before 🤣. Since 💀 is from the *face-fantasy* subcategory with a negative concept and 🤣 is from *face-positive* subcategory with a positive concept, the tweets containing these emojis in this order could be sarcastic or be about attempting to provoke someone. Our analysis of the tweets containing 💀 before 🤣 appear to support the case that such tweets have a jokingly or sarcastic tone.

Another approach to visualize the network is to use nested block model [12]. Figure 3 shows the directed network of emojis for organ donation dataset. In this network, we find out that there are 21 communities of emojis in the first level. Those 21 communities are divided to 8 communities in the higher level. Then they merge to 2 major hyper communities. For each community, we pick some random emojis to show next to their communities.

Later in this paper we will discuss the category-based entropy of communities as a way to understand whether emojis are grouped based on their pre-assigned categories (see Table 1 for some of the categories).

4.2 Network analysis

The first analyses of the networks of emojis is a look at degree distribution. If the distribution can be fitted via a power law function, it means that there are few emojis that have connections with a significant of emojis, while most of the emojis have connections to few emojis. We try to investigate the mentioned distributions in both directed and undirected networks.

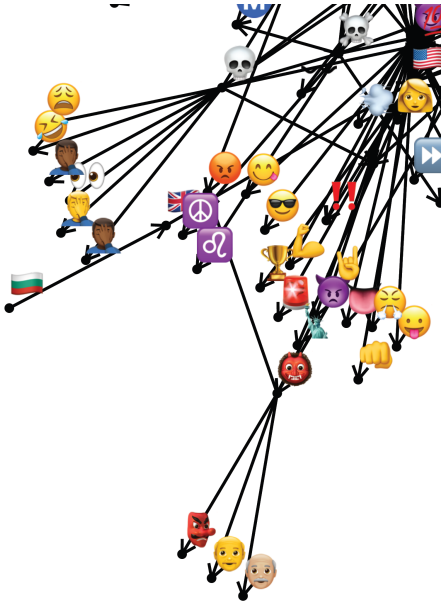

Fig. 2 Part of the directed network of emojis (EmojiNet) for the random (*randSample*) dataset. The edges with the weight less than 8,000 have been deleted for the purpose of visualization.

Along with in-degree distributions, we investigate the edge-weight and weight-degree distributions, because the network is created based on the number of times two emojis accompanied each other in tweets. In reality, the information of usage yields a weighted network. Therefore, the edge-weight distribution is the critical information we can have from these networks because they represent the co-occurrence of pairs of emojis.

We use log-likelihood ratio to find which function best fit the distributions found. A positive result from log likelihood ratio means that the second (right side) function is fitted better on the data. On the contrary, the negative result means that the first (left side) function is more suitable to explain the data. We try 4 distribution functions: power law, truncated power law, exponential, stretched exponential.

Table 3 summarizes our distribution analysis on the emojiNets we created. In the case that log likelihood ratio is close to zero and the p-value is larger than 0.5, we assume that two functions are fitted for the distribution in almost the same level, but we chose the one that is slightly better first. For example, the in-degree distribution for the *Organ* dataset has the truncated power law (TPL) and stretched exponential (SE) almost equally fitted, but TPL is a little better. In all cases the least appropriate function to fit the distributions is an exponential. As it can be seen in Table 3, truncated power law (TPL) is the best function for several of the distributions.

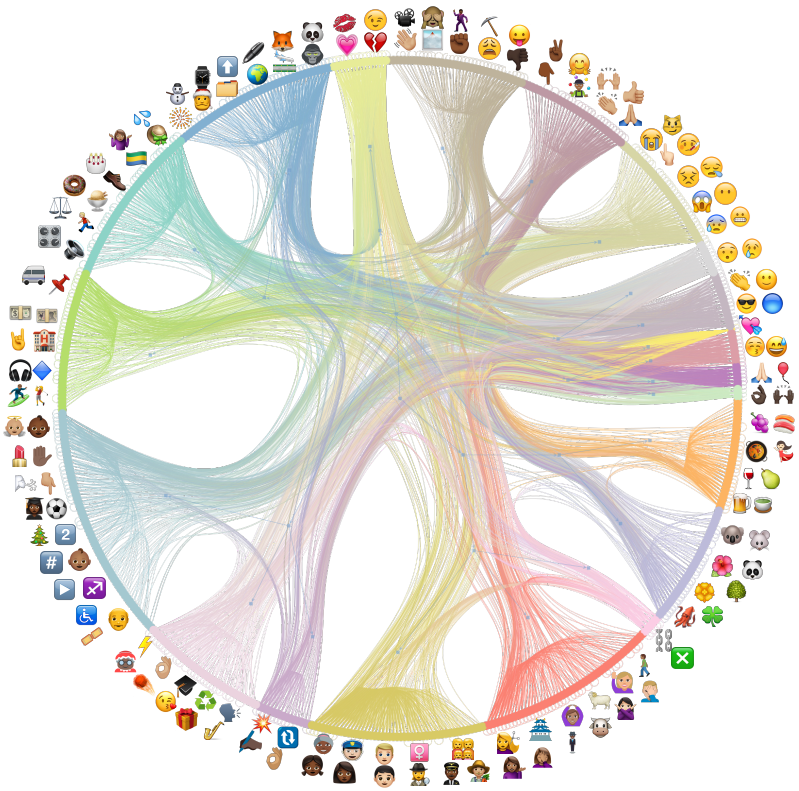

Fig. 3 Nested block model of the communities for *Organ* dataset. In this network, we have 21 communities in the first level, 8 communities in the second level, and 2 communities in the third level. Emojis are picked from each community and put next to the corresponding community as a depiction of what the community represents.

Table 3 The best fitted curve for in-degree and edge-weight for directed EmojiNet; and also weight-degree and edge-weight for undirected distribution analysis based on log-likelihood ratio ($logL$) for all datasets. TPL stands for truncated power law and SE indicates stretched exponential.

Dataset	In-degree	Directed edge-weight	Weight-degree	Edge-weight
G-20	SE & TPL	SE & TPL	TPL	TPL
Organ	TPL & SE	TPL	TPL	TPL
rioSports	TPL	TPL	TPL	TPL
rioTerms	TPL	TPL	TPL	TPL
WWC	TPL	TPL	TPL	TPL
randSample	TPL & SE	TPL	SE & TPL	TPL

After distribution analysis, we looked closer at the directed emoji network to find the important emojis based on different criteria. Table 4 shows the top 5 important

emojis based on three different criteria: frequency of usage, pagerank, and node betweenness. The rightmost column in this table shows the emojis that were considered important in all three main criteria. We want to see the potential differences between various sets of important emojis. The frequency of usage is just the count of the emojis that have been used in the datasets; it tells us how frequent an emoji is selected by users and can also be considered as a popularity metric. Pagerank [10] tells us the importance of an emoji as a function of the importance of other emojis that came before it. For example, if 😂 has a high pagerank, most users use 😂 at the end of their tweets. The node betweenness centrality [9] tells us how much an emoji falls between two emojis; it can give us information about how much an emoji could be seen as a "linking" emoji, perhaps linking ideas expressed by other emojis. Last, we extract the intersection of the set of important emojis to find the important emojis based on all 3 criteria. This set of "common emojis" might represent more versatile emojis because they may be assumed to have different roles. We extract the set of common emojis from top 10 popular emojis and not only the 5 shown in each of the first three columns of Table 4. Let us pick one case of these common emojis, the one related to the *WWC* dataset: ⚽🇺🇸❤️😍🏆. When we verify what happened in that 2015 Women World Cup, we find that the United States won the event. Hence it is not surprising that we have soccer ball, the United state's flag, and a cup in the set of important emojis. The same phenomenon can also be observed from the common emojis of *rioTerms* dataset also. It may support the idea that the common emojis can convey valuable information about the context of the conversation.

Table 4 Popular emojis based on three different criteria. The emojis are ordered from left to right based on their popularity. The common emojis (rightmost column) are the intersection of the top-10 emojis in the three criteria displayed.

Dataset	Frequency	Pagerank	Node betweenness	Common emojis
G-20	😂❤️💕😍😭	❤️💕😂🌸✨	😂😭❤️💩👍	😍💕😂😭❤️
Organ	😢😭❤️🙏😂	❤️🙏😂🖤🤷	❤️😂🙏♀️💕	❤️😂🙏😭
rioSports	👏💚💛😂🥇	👏🥇💚💛❤️	👏💪❤️🥇😂	👏🥇❤️💪😂
rioTerms	🥇😂👏🇺🇸🍗	🥇👏🇺🇸🍗😂	🥇🇺🇸😂❤️👏	🥇🇺🇸😂👏❤️
WWC	🇺🇸⚽👏❤️⚪	🇺🇸⚽⚪🏆❤️	⚽🇺🇸❤️😍🎉	⚽🇺🇸❤️😍🏆
randSample	😂❤️😭😍💕	😂😍❤️😭🔥	♂️♀️❤️😂😠	😂❤️😍✨

In order to have a better insight on how the emojis are used together we extracted communities from the emojiNets. We used a *community* package in Python which implements the Louvain method [3] for community detection. It receives an undirected network and returns the communities. The Louvain method is a greedy algorithm that tries to increase the modularity of the network and starts with finding the small local communities. Figure 4 shows the composition of top 10 largest communities extracted from *Organ* dataset. As it is shown, no community is pure and

they contain emojis from different categories. However, *smiley-people* and *symbols* are common in most communities.

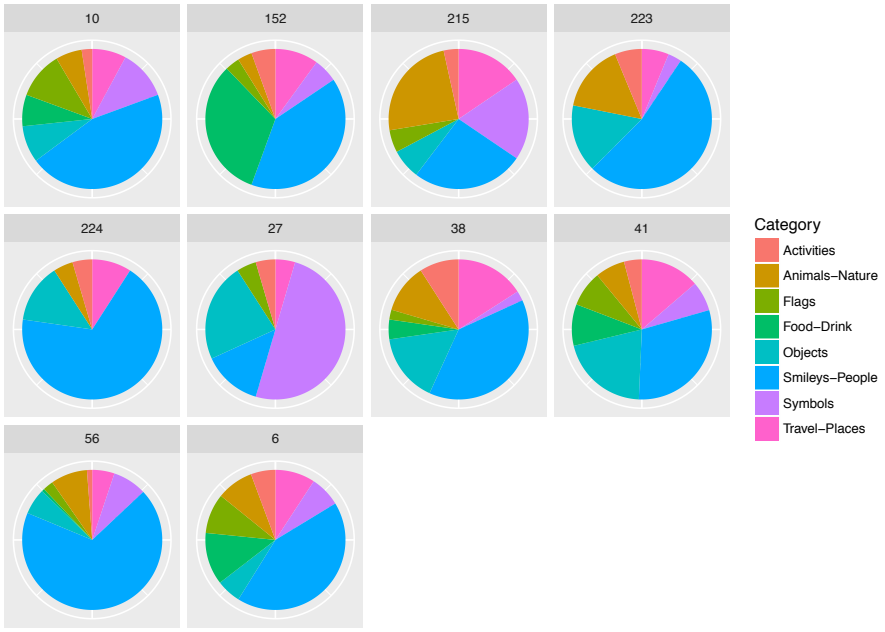

Fig. 4 Top community composition for the *Organ* dataset. The compositions are extracted based on the categories mentioned in Table 1. The numbers on communities are their index for further addressing. Smileys seem to be the most used ones in most of the communities. These results support the claim that emojis from the same category do not form a community.

The entropy of the communities can give us a better information on how emojis are used. We calculate the entropy of the communities with respect to the categories they belong to. Equations 1 and 2 are used to calculate the entropy.

$$e_c = -\sum_{i=1}^{|Orders|} p_{ci} \log_2(p_{ci}), \tag{1}$$

$$p_{ci} = \frac{\textit{\# of emojis in community c from category i}}{\textit{\# of emojis in community c}}, \tag{2}$$

where $Orders$ represent the categories of emojis as in Table 1, e_c is the entropy of community c, and p_{ci} is the probability of having emojis from category i in community c. The higher the entropy is, the less the information it captures with respect of orders/categories. The maximum entropy is $\log_2 n$ where there are n communities or different classes. In our analysis the maximum entropy is $\log_2 8 = 3$ and it is reached in the case that we have the same number of emojis from all categories. The minimum entropy is 0 and it happens if all emojis are from just one category. Then,

the range of the entropy of each community is $0 < e_c < 3$. Figure 5 shows a plot with the variance of entropy of the 10 largest communities in each dataset. It can be noticed that the average of the entropies are close to the mean of the range of the entropy, i.e. around 1.5. *randSample*, *rioTerms*, and *rioSports* have higher variance. More research is required to explain why we observe this difference between datasets.

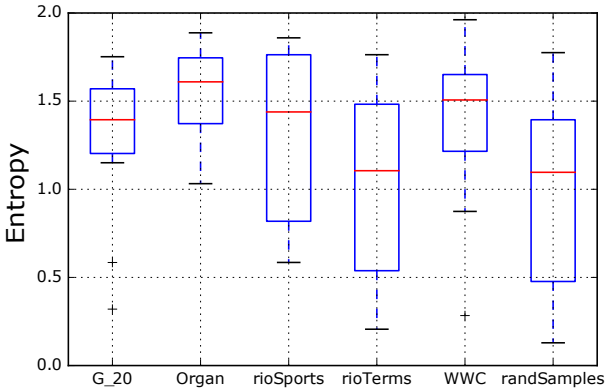

Fig. 5 Entropy of top 10 largest communities by size. The mean entropy for each dataset is close to the half of the maximum entropy.

In addition to entropy of communities, the size of communities may reveal how emojis are grouped. A large community can indicate that a considerable number of emojis have a high chance of being used together. In order to analyze the size of communities, the communities are extracted and then sorted with respect to the number of the emojis they contain. Then, we show the size of the sorted list of top communities in Figure 6. It is interesting that the size of communities drops down after the 4th community and there is no considerable change for community size after that. For example, almost 54% of the emojis are connected to each other in one of the top 4 communities in *randSample* dataset. This fact suggests that most of information is captured by 4 possible groups of emoji.

Last, we want to see the entropy of the top 3 communities for each dataset in Figure 7. As it is shown, the subject-based datasets have high entropy the random sample dataset has lower entropy. Again, this result is surprising and could indicate that groups of emoji are better formed when we look at the emoji used without bias towards specific subjects. However, future work is needed here to understand this finding.

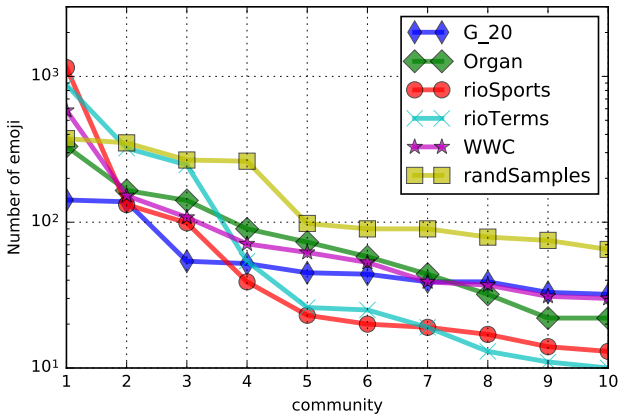

Fig. 6 The size of top 10 communities for all datasets. The size drops down after the 4th largest community.

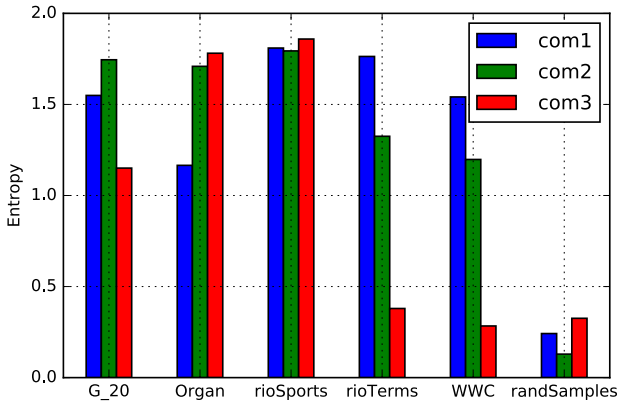

Fig. 7 Entropy of top 3 communities for all datasets.

5 Conclusions and Future Work

In this paper we built a directed network of emojis from 6 datasets collected from Twitter; the datasets represents different areas of interest (subjects) with one being a random sample of tweets. We showed that the distribution of weight-degree and edge-weight follow a truncated power law. Then we showed the important emojis with respect to frequency of usage, Pagerank, and betweenness centrality. We realize that the important emojis are different for each dataset and they are related to the subject of the dataset.

We unveiled the community structure of emojis and they discussed the entropy of top 10 communities. The entropy values for these communities are very similar indicating that people tend to combine emojis from different orders (categories).

As a future work, we will focus on analyzing the entropy distribution of emoji communities using different methods of community detection, in particular we are intersted in using the purity method proposed by Hartman et al [7]. Furthermore, we want to investigate emojis as languages understanding whether we have a semantic structure that could be extracted from the sequence used by individuals.

Acknowledgements We would like to thank the NSF for the grant No. 1560345 that supports this research. We also appreciate the data provided by Diogo Pacheco, Josemar F. da Cruz, and Diego Pinheiro.

References

1. Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
2. Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *LREC*, 2016.
3. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
4. Unicode Consortium. Unicode emoji, 2017.
5. Vyvyan Evans. *The Emoji Code: The Linguistics Behind Smiley Faces and Scaredy Cats*. Picador USA, 2017.
6. Julian Gottke. Instagram emoji study emojis lead to higher interactions, 2017.
7. Ryan Hartman, Josemar Faustino, Diego Pinheiro, and Ronaldo Menezes. Assessing the suitability of network community detection to available meta-data using rank stability. In *Proceedings of the International Conference on Web Intelligence*, pages 162–169. ACM, 2017.
8. Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 770–780. ACM, 2016.
9. Mark Newman. *Networks: an introduction*. Oxford university press, 2010.
10. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.
11. Umashanthi Pavalanathan and Jacob Eisenstein. Emoticons vs. emojis on twitter: A causal inference approach. *arXiv preprint arXiv:1510.08480*, 2015.
12. Tiago P Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, 2014.
13. SM Mahdi Seyednezhad and Ronaldo Menezes. Understanding subject-based emoji usage using network science. In *Workshop on Complex Networks CompleNet*, pages 151–159. Springer, 2017.
14. Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. Emojinet: Building a machine readable sense inventory for emoji. In *International Conference on Social Informatics*, pages 527–541. Springer, 2016.