Understanding Subject-based Emoji Usage using Network Science

S.M. Mahdi Seyednezhad and Ronaldo Menezes

BioComplex Laboratory, School of Computing
Florida Institute of Technology, Melbourne, USA
sseyednezhad2013@my.fit.edu
rmenezes@fit.edu

Abstract. The use of "Emoticons" and "Emojis" in social media as well as most online writing has become the *de-facto* standard on how to express emotions, feelings, etc. Although there are more that 1,000 emojis, not much has been done to understand the way in which people use these characters. The large set of emojis available brings two questions: *(i)* How can users make full use of the emojis available? and *(ii)* Would it be possible to build a recommendation system for emoji usage in text? This paper moves towards a greater understanding of emoji usage by mapping possible relations between these special characters in common text. We look at possible regularities in emoji usages in written, subject-specific, text corpora. We build co-occurrence networks of emoji based on two datasets and show that the structure of these networks are not random and more like a truncated power-law, but more interesting, we show that the structure has similar characteristics despite the text being subject-specific.

Keywords: Emoji, Word Co-occurrence Networks, Network Science, Twitter Data

1 Introduction

Our inability to express emotions in written language is notorious. For instance, who has never tried to send an email with some sarcasm and found that it was not well understood. The misunderstanding arises because the text does not convey your facial expressions or perhaps your tone of voice; crucial for sarcasm. In 1982, Scott Fahlman, a professor at Carnegie Mellon University (CMU) proposed what is considered the first use of a emoticon in a message to a general CMU mailing list.

After the first use, the idea of emoticons spread quickly and many variations have been proposed by Fahlman and others and today emoticons are still commonly used. Emojis were introduced in 2010 in Unicode 6.0 and today there are 1,088 emojis defined in Unicode 9.0. These emojis are graphical version of the emoticons and include representations such as 👕, 🍺, 💃, and 😊. With the growth in the number of Internet users, the need for the emojis has been risen. The variety of emojis correspond to the diversity of emotional feelings in humans [8] but it also grew to other usages such as flags, animals, symbols, activities, etc.

Despite their popularity (e.g. emoji are used in nearly 800% more campaigns than in 2015[1]), there has been little movement on trying to understand how society uses these emojis. Even though,"emojis won the battle of words" as claimed by the New York Times [2], their use relies completely on user knowledge about a particular instance of the characters. The popularity of emojis has lead the Oxford dictionary to select the word of 2015 as "face with tears of joy" which is the name of an emoji (😂).

Another interesting aspect about the emoji phenomena is that they become akin to a universal language because many are understood similarly in different locations easing the connection of people from different cultures [6]. As a matter of fact, emojis can be useful tools to analyze social media because first, they are widely used by people from different countries and second, they have been adopted in different social media, such as Facebook, Twitter, etc. Furthermore, they are employed for purposes other than social media, such as mobile phone notification using emojis [10]. On the other hand, some emojis are ambiguous in their meaning leading to different usages. One of the most common cases is the "Person With Folded Hands" (🙏), which in some cultures (such as in Japan) is seen as "please" or "thank you", while in others (such as in Brazil) is widely used as a sign for prayer or "amem".

2 Related works

One of the works to help computers understand emojis, attempts to build an inventory of meanings for emojis in a way that is easy for machines to understand. Wijeratne et al. [11] tried to a make connection between each emoji and its meaning in words. The output of their work is a semantic network in BabelNet. Although they try to have a comprehensive machine readable network of emojis and words, it could have been better if they considered the co-occurrence of emojis in social media with other frequent words and have an analysis on their bipartite notwork of words and emojis. Besides, a combination of emoji sentiment analysis [7] with words may give us a more accurate list of emoji meanings. In [1] a vector space model has been used for Twitter data in order to connects emojis to meaningful corresponding words.

The number of emojis that are being used in Twitter can be found on emojitracker.com. Furthermore, in [9], the authors discuss social aspects related to emoji usage; they argue that Twitter users who embrace emojis tend to keep using them instead of emoticons, thus the number of emoticons being used is falling down. The study on emoji usage has also been done in a geocentric way. Scholars focused on the emoji distribution both over the world and in countries. For instance, Ljubešic and Fišer [5], gathered information about emoji usage distribution by country and investigated the emoji popularity for the whole world in this geocentric approach. Then found the list of popular emojis for each region, followed by a clustering of the countries based on emoji popularity, they found that countries could be classified into four different group based on the "most distinctive emojis". Finally, they discovered a correlation between some emojis and some world development indicators of the world bank. For example, surprisingly, countries

[1] https://www.appboy.com/blog/emojis-used-in-777-more-campaigns/
[2] http://www.nytimes.com/2014/07/27/fashion/emoji-have-won-the-battle-of-words.html

with high life expectancy use “face with tears of joy” (😂) less often than the countries with low life expectancy.

As we mentioned before, having a network of emojis based on their co-occurrence may help us analyzing emojis from a different angle. Lu et al. [6] concentrated on trying to understand human behavior in the context of culture from data gathered from users of smartphones. Accordingly, the authors correlate the culture index with emoji sentiments. They considered the cultural index introduced by Hofstede in [4] that delineated the social differences with six features. For example, power distance is one of them. This feature expresses how much people with less power accept that power is distributed unevenly. They discovered that strong power-distance countries use more negative emotions with emojis.

3 Data Handling and Network Extraction

Ths is based on the subjects of tweets, we define “subjects” as the theme used for the collection of theses datasets. In this initial work, we selected two diametrically different datasets in order to verify possible structural differences. Recall that our approach argues that the structure may be linked to the subject of the conversation. We created two emoji networks, one for each dataset. The list of emojis used here are from apps.timwhitlock.info.

The datasets were named *WWC* and *ProgLang*. For *WWC*, the tweets were collected during the 1-month period of the Women World Cup and Americas Cup (soccer) held in the USA in June 2015. This dataset contains more than 10 million individual tweets. The *ProgLang* dataset contains tweets from September the 20th to November the 1st in 2016 related to computer programming languages. The dataset contains approximately 2.5 million tweets.

3.1 Building an Emoji Network from Tweets

The process of creating an emoji network from tweets is quite straightforward. The general idea consists of sifting through each tweet and looking for emojis in the dataset. Each tweet generates a k-clique where k is the number of emojis in that tweet; Fig. 1 shows the process in which networks are extracted from the unstructured textual data.

Following the generation of these k-cliques, they are combined forming a larger network. The resulting network for the example shown in Fig. 1 is depicted in Fig. 2. Note that the network is weighted and the edge (👍,😍) has weight 2 because it appears in Fig. 1 in two different tweets.

4 Experimental Results

After creating the networks of emojis, we performed an analysis of their structure using using Network Science concepts, such as degree distribution, edge-weight distribution, network density, to name a few [3]. We start with general network characterization but we also discuss similarities and differences between the two networks we worked on.

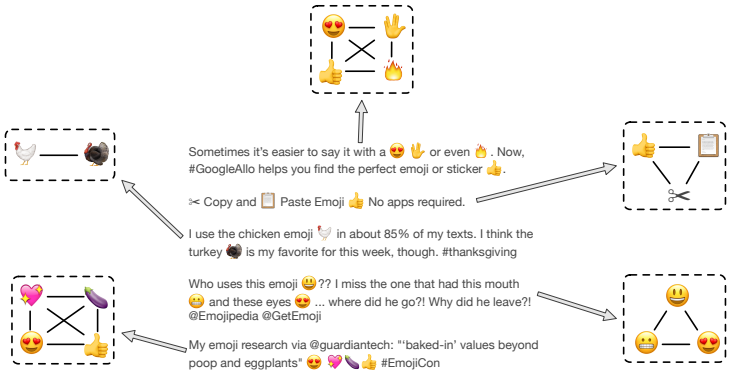

Fig. 1. Process describing the extraction of a k-clique from each individual tweets, where k is the number of emojis in each tweet. The network is undirected.

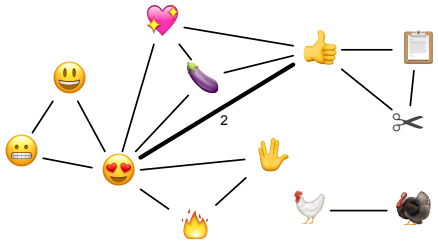

Fig. 2. Result of the combination of all the k-cliques in Fig. 1. This network is undirected and weighted.

4.1 Network Characteristics

There are several important characteristics that can be extracted from networks. Table 1 shows some basic network properties for both networks. It also shows the result of three important aspects in these networks, **Node Betweenness**: A node has high betweenness if it happens to frequently be in the shortest path between other pair of nodes. **Edge Betweenness**: An edge has high betweenness if it happens to frequently be in the shortest path between pair of nodes. **Assortativity**: It is a measure of how often a node with a particular degree connects to others of similar degree. High assortativity means that nodes connect to others alike; the metric assumes values between -1 and $+1$ for dissortativity and assortativity respectively [12].

We calculated the assortativity of the network; both networks are slightly disassortative meaning that the nodes with higher degree tends to have connections with nodes with lower degree. The *WWC* network is more dissortative.

For the analysis of betweenness, Table 1 shows the grinning face (😬) has the highest node betweenness in both datasets confirming the popularity of this emoji regardless of the subject area. Another interesting result from Table 1 is the fact that the maxi-

Table 1. Properties of the two networks used in this study. The *WWC* network is a lot more dissortative, while the programming languages is neutral.

Dataset	Max Weight	Average Weight	Max Node Betweenness	Max Edge Betweenness	Assortativity
ProgLang	510	3.73	😁 & 🍃	(❤️,🕑)	-0.066
WWC	71,099	34.08	😁 & 🌵	(😏,🈷️)	-0.193

mum edge betweeeness occurs for the edge linking the smirking face (😏) and squared Chinese-Japanese-Korean character (🈷️). It is amusing because smirking face is one of the top favorite emojis in the United States [6], the squared cjk is related to Japanese characters, and the final of the 2015 world cup was USA against Japan. It appears then that if one knows the semantics of these emojis, it may be possible to learn something about the subject area from which they were extracted and this indications opens a door for possible recommendation systems.

4.2 Degree and Weight Degree Distributions

One of the most common characteristics scientists measure in a weighted network are both the degree and weighted degree distributions of nodes [2]. We tried to fit common functions found in real-world networks and used log-likelihood ratio—denoted by $L(d_1, d_2)$—for distribution analysis. The positive values of log-likelihood tell us that the left function d_1 is a better fit to the original data, and d_2 otherwise.

In Fig. 3 we demonstrate several possible fitted functions for degree and weighted degree distributions of the *WWC* data set. The red dotted line show the data and the lines are the functions that could possibly fit the data distribution. A visual inspection can immediately say that the exponential function is not a good fit for weight-degree distribution. For a more complete analysis of the goodness of fit, we show in Table 2 the log-likelihood ratio and the p-value between different functions with respect to *WWC* data set. The results show us that stretched exponential is the best fit function for degrees.

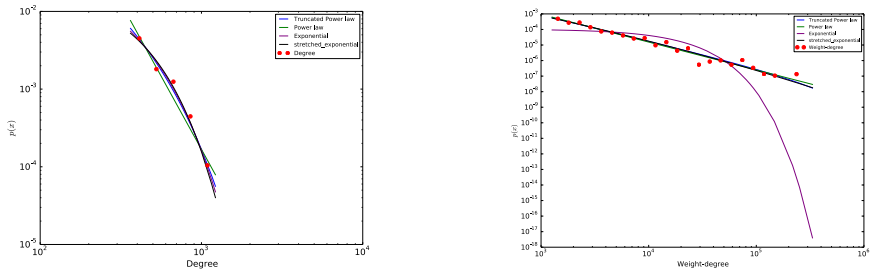

Fig. 3. Fitting applied to the degree and weighted-degree distributions for *WWC*.

Table 2. Log-likelihood ratio ($logL$) for degree, weighted degree and edge-weight distributions for the *WWC* network.

Functions	Degree		Weighted Degree		Edge Weight	
(d_1, d_2)	$logL$	p-value	$logL$	p-value	$logL$	p-value
(powerlaw, exponential)	-3.96	0.015	173.35	0.000	878.31	0.000
(powerlaw, truncated power-law)	-3.59	0.007	-2.77	0.018	-5.01	0.001
(powerlaw, stretched exponential)	-4.25	0.063	-1.98	0.243	-2.44	0.544
(truncated power-law, exponential)	-0.37	0.289	176.13	0.000	883.32	0.000
(truncated power-law, stretched exponential)	-0.66	0.509	0.79	0.263	2.57	0.279
(exponential, stretched exponential)	-0.29	0.444	-175.34	0.000	-880.74	0.000

In addition to the degree analysis, another important aspect of our emoji network are the weights of edges. The edge weight represents how pronounced the co-occurrence of pairs of emojis are in the dataset. Hence it is important to characterize this distribution to understand how the values of edges are distributed. Fig. 4 shows the fitting of the edge-weight distribution for *WWC* network. We also preformed a log-likelihood analysis and found that the best fitted distribution of this is a truncated power-law. This means that there are relatively fewer pair of emojis that are popular and that most pairs are rare.

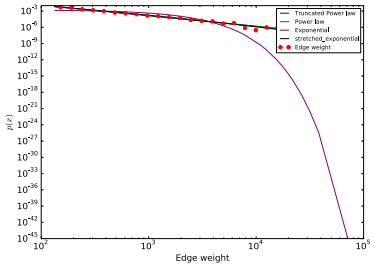

Fig. 4. Edge-weight distribution for the *WWC* Network.

In this paper we also reconstructed an emoji network from another dataset related to programming languages. Similar to what we have done for the *WWC* network, we analyzed the network degree and weighted degree distributions, as well as the edge-weight distribution. The degree distributions are depicted in Fig. 5.

Furthermore, Fig. 6 shows the best fitted function for the edge-weight distribution as being a truncated power-law which again agrees with what was found for the *WWC* network.

The fitting of the functions was done again using an approach based on the log-likelihood ratio. In Table 3 we find more details about the pairwise comparison between different functions for degree, weighted-degree, and edge-weight. As one can observe, the best fitted function favors a stretched exponential for degrees, while for weighted-degree and edge-weight, the truncated power law is clearly the best fit.

In summary, in both data sets, we have the same type of distribution for degree, weighted-degree and edge-weight values of the networks. This is a preliminary work but

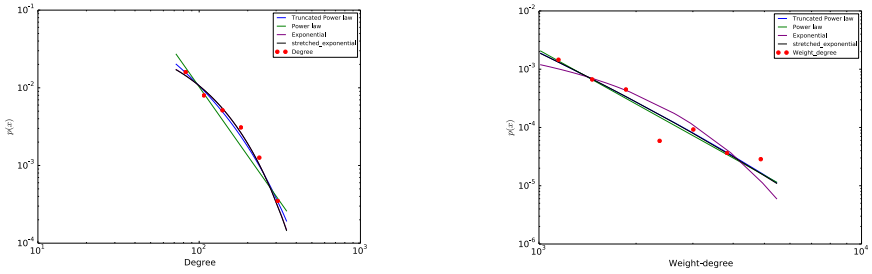

Fig. 5. Fitting applied to the degree and weighted-degree distributions for the *ProgLang* network.

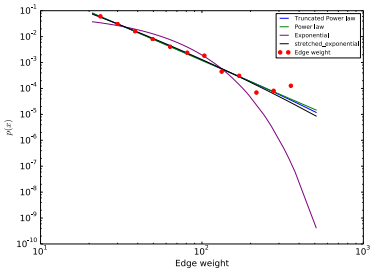

Fig. 6. Edge-weight distribution for the *ProgLang* Network.

it does seem to indicate that structure of emoji usage is not much affected by the subject of the conversation. Note that this does not mean that the emojis used are the same, quite the contrary, our work only argues that the networks formed from the co-occurrence have similar structures but it is very likely that different emojis occupy similar structural positions in the different networks. Table 1 supports this claim in our two datasets.

5 Conclusion

In this paper we constructed co-occurrence networks from emojis and analyzed their structure to understand possible regularities. We used two datasets and showed that

Table 3. Log-likelihood ratio ($logL$) for degree, weighted degree and edge-weight distribution for the *ProgLang* network.

Functions	Degree		Weighted Degree		Edge Weight	
(d_1, d_2)	$logL$	p-value	$logL$	p-value	$logL$	p-value
(powerlaw, exponential)	-11.08	0.002	1.72	0.347	100.04	0.000
(powerlaw, truncated power-law)	-10.30	0.000	-0.17	0.556	-0.52	0.310
(powerlaw, stretched exponential)	-11.08	0.002	-0.09	0.782	1.54	0.397
(truncated power-law, exponential)	-0.77	0.499	1.92	0.215	100.56	0.000
(truncated power-law, stretched exponential)	-0.78	0.525	0.08	0.206	2.07	0.105
(exponential, stretched exponential)	0.01	0.922	-1.84	0.055	-98.48	0.000

although they do not seem to have a structure similar to network of words in written language or other common real-world networks, they do have similar structures among the two datasets.

We are working on larger datasets. In these, we will focus on community detection as a way to find family of emojis and whether the families correlate to classes of emoji (flags, professions, etc.) Furthermore, PageRank could be useful to understand the importance of emojis to language; for this we need to have a directed version and we are investigating if the order they appear in the text could realistically represent a direction. For instance if one writes "I ❤️ to have a 🍺" or something such as "🍺 is one of the things I ❤️" have slightly different meanings due to the order the emojis are used but also the relation to the words in the sentence. A directed network of usage could capture some of these nuances.

References

1. Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *Language Resources and Evaluation conference, LREC, Portoroz, Slovenia*, 2016.
2. Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.
3. Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *science*, 323(5916):892–895, 2009.
4. Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and organizations: Software of the mind*, volume 2. Citeseer, 1991.
5. Nikola Ljubešic and Darja Fišer. A global analysis of emoji usage. *ACL 2016*, page 82, 2016.
6. 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.
7. Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.
8. Jaak Panksepp. Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and cognition*, 14(1):30–80, 2005.
9. Umashanthi Pavalanathan and Jacob Eisenstein. Emoticons vs. emojis on twitter: A causal inference approach. *arXiv preprint arXiv:1510.08480*, 2015.
10. Channary Tauch and Eiman Kanjo. The roles of emojis in mobile phone notifications. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1560–1565. ACM, 2016.
11. 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.
12. Ramón Xulvi-Brunet and Igor M Sokolov. Changing correlations in networks: assortativity and dissortativity. *Acta Physica Polonica B*, 36:1431, 2005.