Analysis of Brand Co-Mentions in Twitter-Tweets

Project - Complex Networks and Web

MSc Web Science and Big Data Analytics

Abstract
Most twitter analytics companies focus on audience benchmarks and audience analytics. However Twitter offers almost real time access to public tweets and hence access to conversations and opinions about brands. This project report was intended as prove of concept for what companies can learn from brand conversations, in particular from brand co-mentions in tweets. During this study various data sources were evaluated. Eventually the twitter search API was used to extract 3.7 million tweets during Christmas 2014. 114.000 of those included brand co-mentions. Those were connected as a network and analysed in respect of it's network properties as well as of it's insightfulness and application for management. The results reveal industry structures and validate relationships between brands and partners e.g. companies can use this to validate their aspirational competitors with the actual ones. Finally this project indicated that such analyses are insightful and it provided inspiration for further research in this area.

Introduction

Companies like PeerIndex and Brandwatch have proved that corporate companies are interested in measuring their social media footprints. They do this because they can gain valuable insights for customer relations and brand management. One goal of this project was to familiarize myself with twitter and twitter data collection while studying the network of brand co-mentions in tweets. In this approach the network was designed as an undirected graph, where the edges connect all distinguishable brands mentioned in the same tweet. Although multiple mentions of the same brand in the same tweet were counted there were no connections created for this cases (no self-loops). Brand-mentions are extracted by looking for words, mentions and hashtags that distinctively represent a brand. A full list of keywords, mentions and hashtags that were used can be found in the attachments[1]. In this report the terms co-mentions and co-occurrences are used interchangeably.

The main goal of this study is to validate that this particular type of network analysis has value and yields valuable insights for brand management. It's intended to be a proof of concept for further research studies in this direction.

Research question(s)

The research questions were:

- What structure do brand Co-Mentions on twitter follow?
- What interesting properties does the network show?
- What are the most influential brands?
- What implications for brand owners can be derived from the insights?

Data and methodology

Data Source Evaluation

In order to make the results meaningful the data should include at least 100.000 tweets including 2+ brands for statistical significance. Moreover tweets should be distinct as far as possible. As this initial analysis is a proof of concept, geography doesn't matter as long tweets are in english (lots of data had to be collected within a short period of time).

During this project the following tweet data sources were evaluated in chronological order:

[1] See Appendix for detailed description of the project file structure

1. Stanford Network Analysis Project
   A general purpose network analysis and graph mining library that used to have a data set with 476 million tweets but that set was removed as per request of twitter[2].
2. sentiment140.com
   An online service that offers users the ability to run twitter sentiment analysis for custom queries and provides a twitter tweets dataset incl. sentiment for students with 1.6 million tweets. I started using this set however there were only 700 tweets where 2 or more brands co-occurred.[3]
3. Twitter Streaming API
   As the name implies this is an API service that streams twitter tweets as they come. There are ways to filter tweets that include certain words. I initially started to go down this route however there were a couple of challenges. For instance not all brands are getting a comparable amount of tweets in the same time. Secondly there was a huge amount of spam that would have required to be dealt with.[4]
4. Twitter REST API
   This provides endpoints for tweet searches, user timelines and many more properties. The downside is that those endpoints have limitations for how many API calls could be made in 15 minute windows.[5]

Eventually i decided the best fit was the Twitter REST API. Even though it was more work to collect and process the data. Despite the limitations of the REST API it provided the best control and data quality for the purpose of this project.

Data Processing

In order to access the API, I used python and tweepy.[6]. The list of brands was manually created by walking through fanpagelist.com[7] and wikipedia[8] as well as adding missing well known brands manually. The initial list included 240 brands. For every brand I added twitter profile names, hashtags and name variations of the brands as keywords e.g. BA for British Airways or VW for Volkswagen, etc.. The list of brands including all used keywords can be found in a json file in the attachments[9].

The approach to pull the branded terms was iterative. For each brand the last 1000 branded english tweets were pulled with the result-type “mixed”[10]. This option ensured that the response included both popular and real time results[11]. In the next step out of the 1000 branded tweet results the program randomly selected 100 brand search results. Next from the

[2] http://snap.stanford.edu/data/twitter7.html

[3] http://help.sentiment140.com/for-students/

[4] https://dev.twitter.com/streaming/overview

[5] https://dev.twitter.com/rest/public

[6] https://github.com/tweepy/tweepy

[7] http://fanpagelist.com/ - website that lists the top followed and liked brands on twitter and facebook

[8] http://en.wikipedia.org/wiki/Category:Lists_of_brands

[9] Brands path = ./data/brand_keywords.json

[10] I had to perform 10 queries per brand as the search endpoint only returns 100 tweets per search. In order to not download the same results again I also had to use paging with max_id

[11] https://dev.twitter.com/rest/reference/get/search/tweets

timelines of the users who tweeted it, the last 200 tweets were downloaded and used. It took 20h to download all the timelines (240 brands * 100 twitter user timelines / (300 API calls limit per 15 minutes) / 60 minutes). Only the results from the user timelines were taken into account. Using the user data also allowed me to perform an analysis of brand co-mentions in the users timelines. Finally the tweets were tokenized as uni-, bi- and trigrams (in lower case ignoring special characters) as some brand names consist of two or three words. Finally more brands that occurred in high frequency in the dataset were added so the parser would recognized those brands too.

Tools

Gephi: For visualization of the network and calculation of the network properties

R: For calculation and plotting of the network statistics

NetworkX: For calculating the Rich-Club-Coefficient as Gephi doesn't provide this metric

Results

For this study 3,788,156 tweets from 21,030[12] users were captured from Dec 21 to Dec 24 2014. Chart 1 shows that in 86% of tweets there are no brand mentions. 11% of all tweets included only a single brand which left with 3.1% of all collected tweets with two or more brands mentioned. This resulted in 124,000 co-mentions (edges) extracted out of 118,000 tweets.

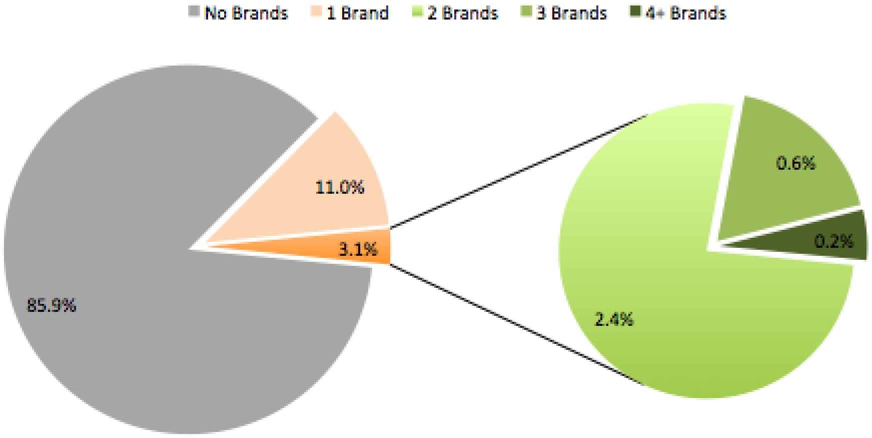

Chart 1: Distribution of Brands mentioned in Tweets

[12] the number of tweets does not equal to 24,000 as some brand queries returned less than 100 results

In the following sections we're going to start looking at the results in a qualitative-visual way and then look at the quantitative network properties such as degree distribution, betweenness, closeness and more.

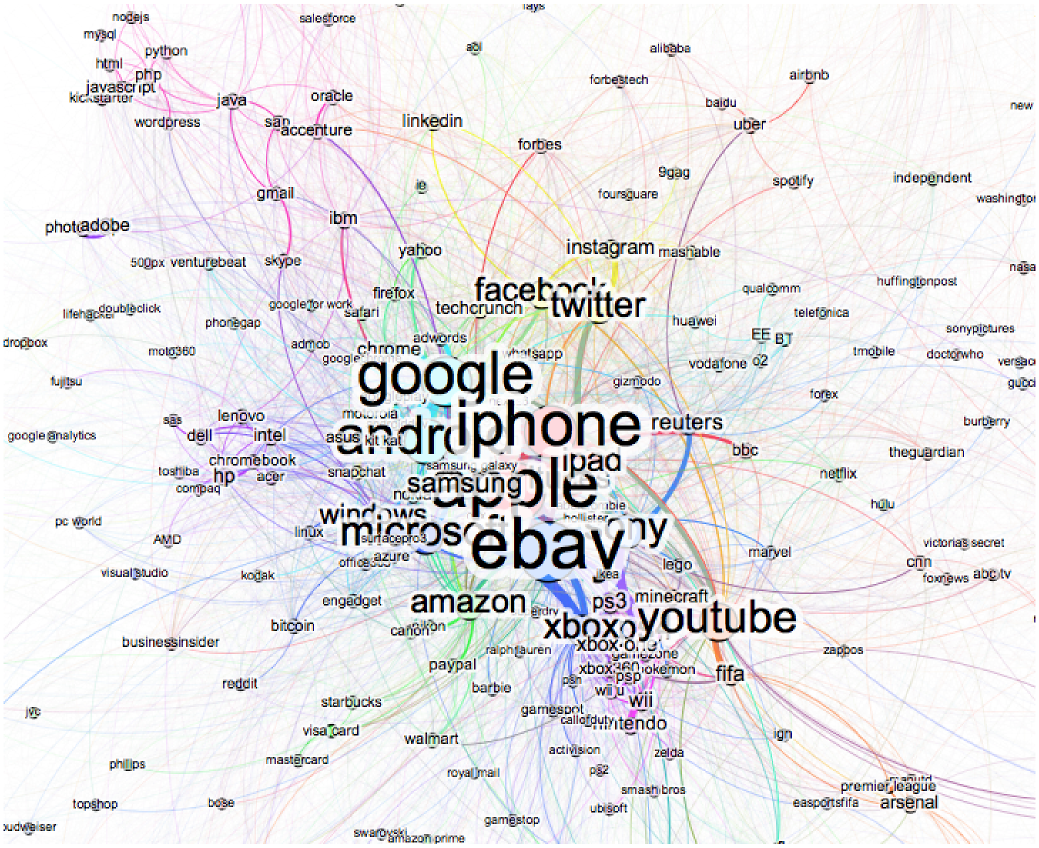

Chart 2: Brand Co-Mentions Network

Brand Co-Mentions Network

Chart 2 shows an extract of the brand co-mentions network visualized with force atlas 2 where the size of the text indicates a high PageRank and the different colors indicate modularity classes[13]. The graphic shows distinguishable structures between brands. It strikes to the eye that technology/internet brands such as eBay, YouTube, Google, Apple, Amazon and Microsoft are in the center. This is because they all are very well connected between each other as well as to many other brands in the network. On a second look there are a couple of industry clusters visible such as gaming and video consoles, consumer electronics, airlines, fashion, entertainment and finance. For instance when going counterclockwise around the network core starting at the top we'll find a pink colored enterprise technology cluster including IBM, SAP, Oracle and Accenture which are well connected with technologies and programming languages. Next left to the core in purple we can see computer

[13] There were filters applied to the visualization. Edges are only shown if there are at least 2 co-mentions with another brand as well as if a node has at least a degree of 3

manufacturer such as HP or lenovo accompanied by operating systems and hardware manufacturers like intel. A high resolution picture can be found in the attachments. In the next few pages we're going through the qualitative properties of the network.

Degree

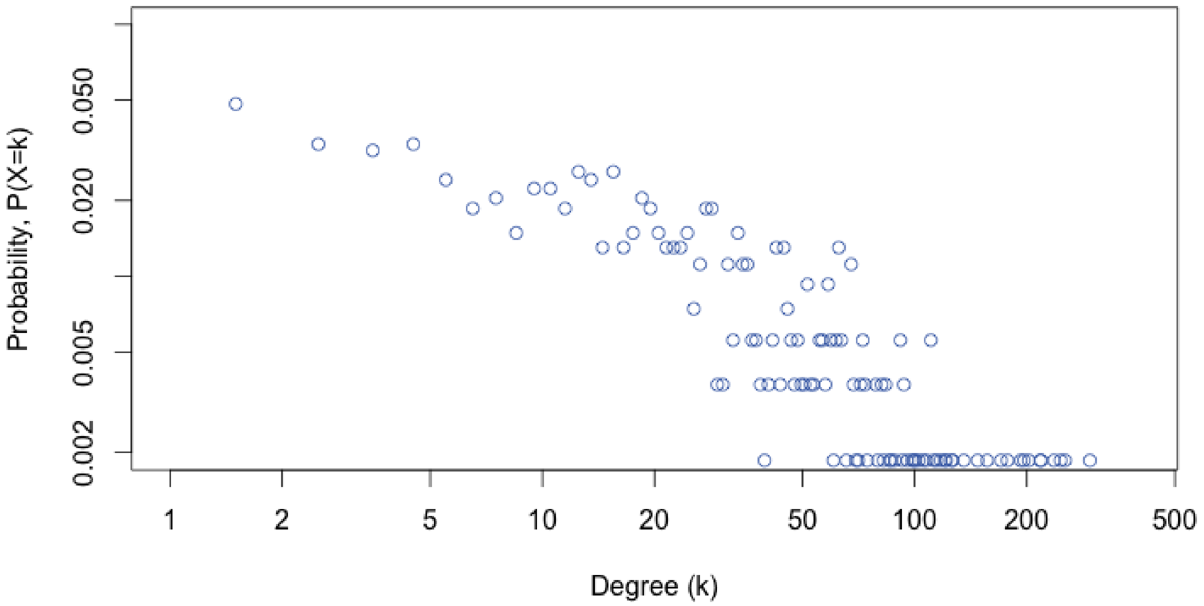

Chart 3: Probability Distribution

It was already visible in the Brand Co-Mention network image that brands like eBay, YouTube, Facebook and Twitter are well connected to each other and have edges to plenty of other brands. This is confirmed by the probability of degrees seen in chart 3 above. It shows that with 50% probability a brand is connected to less than 2 brands. The data shows that 80% of all brands are only connected to 51 brands or less whereas 1% of richest brands (in terms of co-mentions) are connected to over 210 brands and very well interconnected to each other.

Rich-Club Coefficient

Chart 4 shows the Rich-Club Coefficient vs Node Degree which confirms that brands tend to connect with other well connected brands. The graph shows a similar trajectory for the Rich-Club-Coefficient as the www network does. Like the web, where every website can link to another website without any barriers, users can mention (connect) two or more brands together in a single tweet in anyway they want to. An overview of the top twenty best connected brands can be found in the appendix.

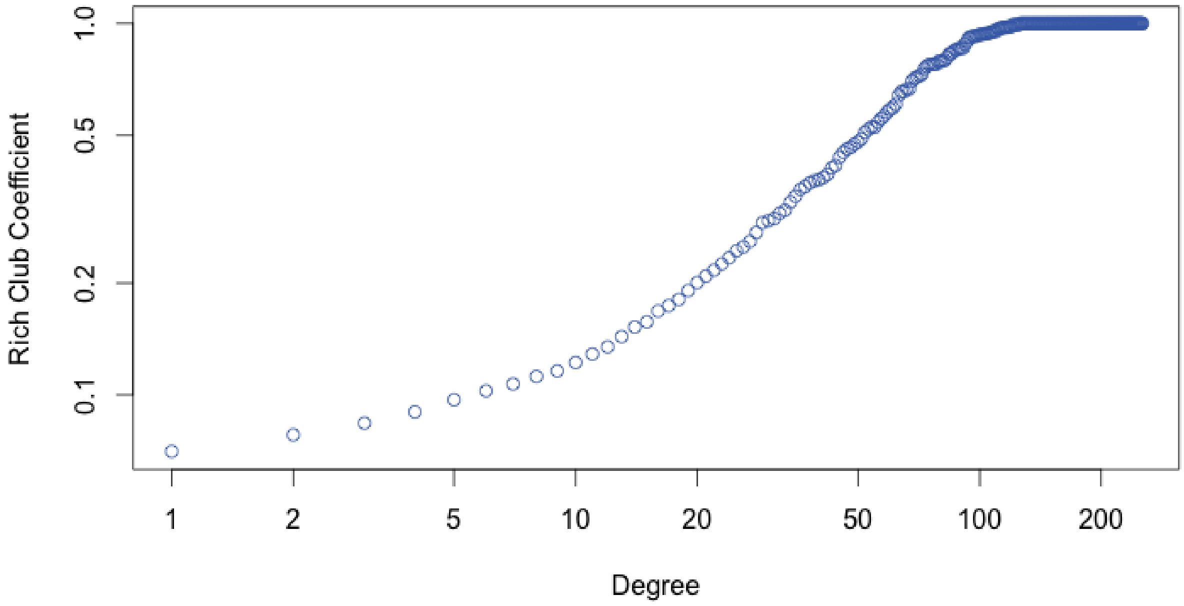

Chart 4: Rich-Club Coefficient vs Node Degree

Betweenness

The comparison of Degree and Betweenness Centrality in chart 5 below illustrates the influence of the highest connected brands. Most brands have a very low betweenness and play a rather unimportant role in the overall network whereas the richest brands are connected to almost all others. However when looking more thoroughly at the chart we can see outliers in between that have slightly higher betweenness centrality than their peers with a similar degree.

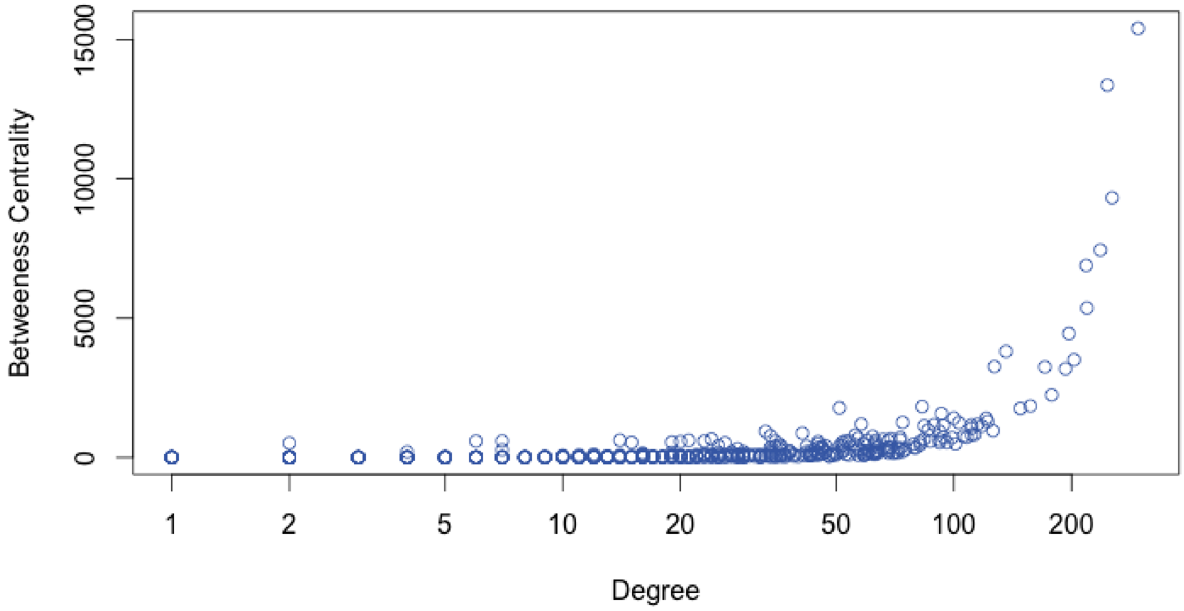

Chart 5: Betweenness Centrality vs Degree

To give few examples of those outliers: The brand Fifa is a bridge for many sport companies but also for airlines like Qatar Air, Emirates airline and Etihad. It is interesting that those particular airlines sponsor many football teams across all leagues in Europe. Adidas and Nike for instance are bridges for a lot of fashion brands. My hypothesis for these results is that the common denominator across fashion oriented people is sport. It doesn't matter what clothing style, and therefore brands, individuals prefer. They all partake in sports and therefore trust the main sports brands. Another type of brands that show high betweenness are market leading brands e.g. BMW in automotive, Samsung in CE or IBM in technology. Those brands tend to have the highest scores for the betweenness metric in their industry. Lastly, and not surprisingly, blogs and news magazines tend to have high betweenness for single or sometimes multiple industries e.g. BBC or TechCrunch.

Closeness Centrality

Closeness centrality is an indicator, in the case of this study, for brands that are highly connected to other brands in their industry (club/cluster). Chart 6 shows the cummulative distribution of closeness centrality in the Brand Co-Mention network.

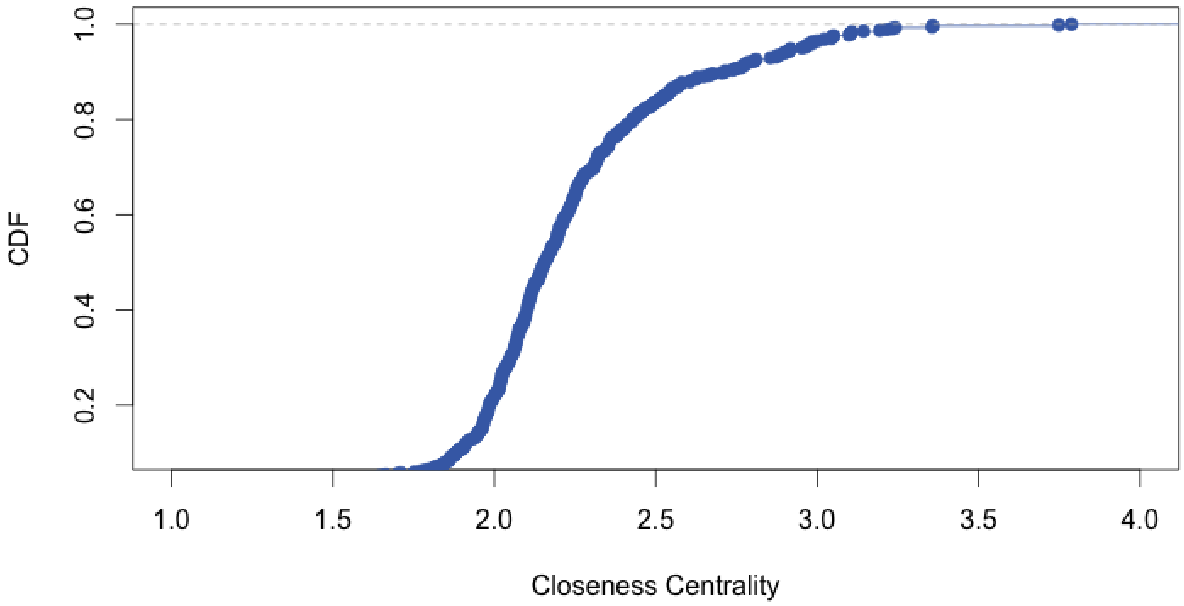

Chart 6: Closeness Centrality Distribution

This metric indicates the average distance from one brand to every other brand in the Brand Co-Mention network. In this study brands with the lowest, and therefore most significant, closeness values are the brands who also have the highest betweenness and degree values. Given the overall centrality as well as the already discovered properties of the network, like rich-club and the characteristic of betweenness centrality, that is no surprise. The network is well connected through the central hubs, chart 6 shows that 80% of brands are only 2.4 brand-mentions away from all other brands which can be seen in the distribution of closeness centrality chart below.

Single Brand Tweets vs. Tweets with more than one brand

Chart 7 below shows a correlation between the number of connections from one brand to other brands with the total number of tweets. The key result is the fewer brands are connected (or better associated) with other brands then the lower the overall number of mentions whereas the higher the degree and association the more mentions they get.

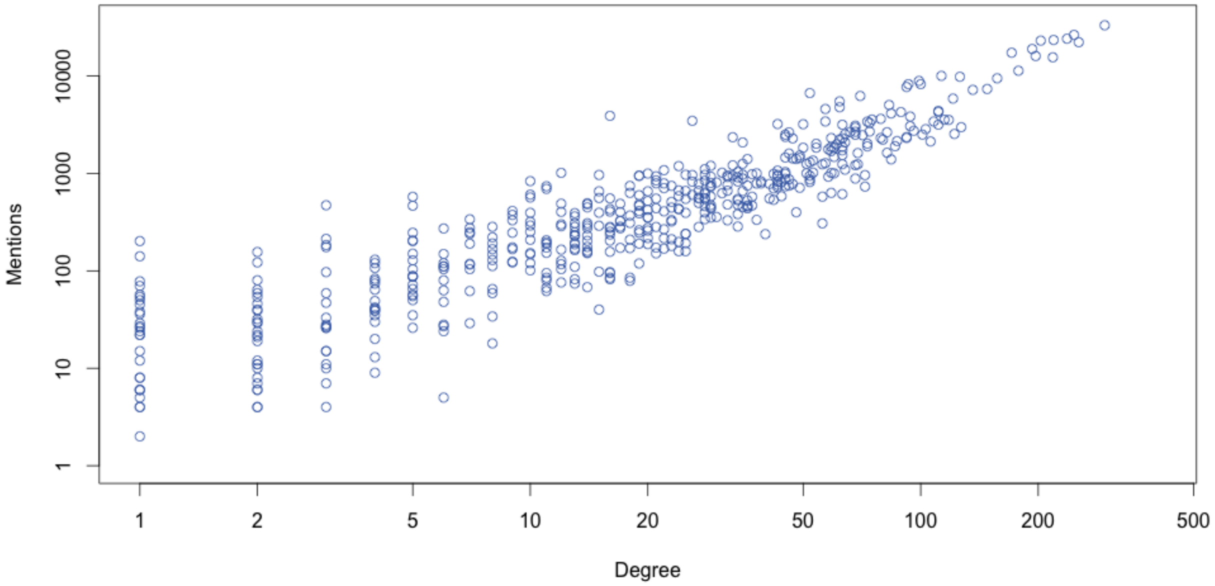

Chart 7: Relationship between Mentions and Degree

Discussion

What brands can learn from Co-Mentions in tweets

There are a couple of analysis which focus on the brand influence or on identifying influencers within the follower network. Most of the times this has the goal to amplify the brand's audience reach. This analysis usually help brands to understand with whom they most efficiently gain influence. The analysis of co-mentions on the other hand has a different purpose and can be looked at from different angles. In the approach introduced, connections between brands were only created when they appeared together in the same tweet. The causes for brands appearing together vary.

In the results we've seen high centrality for a handful of brands. The reason for this high centrality in the Brand Co-Mention network is due to a few hubs with high reach in terms of users outside of twitter. This fact is only indirect reflected in this network. Brands are using hubs like YouTube or Instagram as a platform to carry videos or pictures about their brand or promoting products and distributing the information on twitter. Another example are product brands that get sold at ebay or amazon.

However there are also other cases where brands don't interact with each other because one brand is a specialized platform for content or commerce. To name a few cases:

- user compares brands e.g. likes one over the other, doesn't like them or like them all
- user ask for advice which brand to buy for a specific product category
- user describes a situation involving multiple brands e.g. Spilled Coke in my new Audi
- brands organically relate to each other e.g. apple and iphone or intel and hp
- brand has a sponsorships or a promotion with another brand e.g. mastercard and uefa champions league or qatar airways and FC Barcelona
- or it could be spam where one brand tries to exploit the popularity of another brand

Everything just mentioned, except spam, are genuine brand interactions from users that allow conclusions about the relationship of brands as well as about users relevant brand set and preferences.

As seen in the network visualization the co-mentions uncover industry structures and relationships between brands. One way to look at this connections between brands is looking for clubs (well connected brand-clusters). A brand might think it is or is not part of a club e.g. premium airline club or luxury car manufacturer club. Brands can use this to validate their own perception who their competitors are vs the ones they actually getting associated with. For instance, surprisingly BenQ is not part of the CE and PC manufacturer club. Another example is when we look at the mobile phone industry we can assume that Samsung likes to see themselves competing with Apple and iPhone. Chart 8 shows which of the biggest mobile manufacturer brands are co-mentioned with Samsung. It confirms that Apple is the biggest competitor but also that LG, Nokia and Sony are strong too. Such insights for BenQ or Samsung can validate the competitive position and gives direction for further investigations so a they can take action.

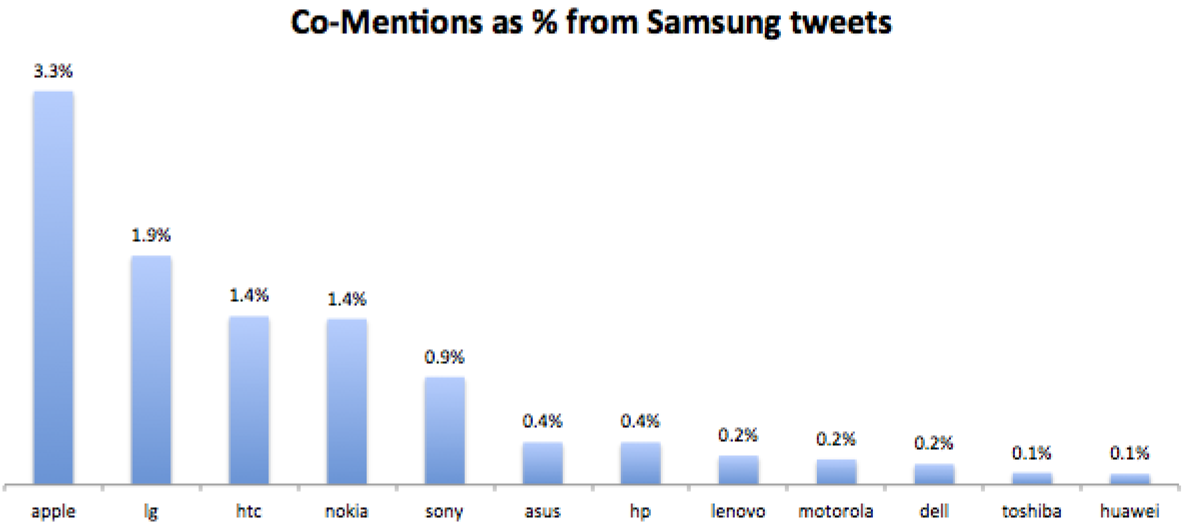

Chart 8: Co-Mentions as % from Samsung tweets

Another insight which can be gained through this analysis is validating if a partnership or sponsorship has an impact for a brand. One example that indicates that a partnership has an effect is the sponsorship of Emirates Airlines of Arsenal London. Emirates has many co-mentions with Arsenal and the Premier League. Hence Emirates are associated with the fifa and football in general.

Further Research Questions

In the analysis shown, the sentiment of users when talking about brands was not taken into account. Nonetheless this could be important to grasp the context of users talking about brands e.g. which brands are mentioned together in a positive and which in a negative way. The resulting graph would be directed and have two different connection types, a positive and a negative one. This would give insights about brand preferences for instance what users prefer one brand over the other. e.g. Apple users doesn't like Samsung. Monitoring changes over time could possibly be used to predict shifts from one brand to another.

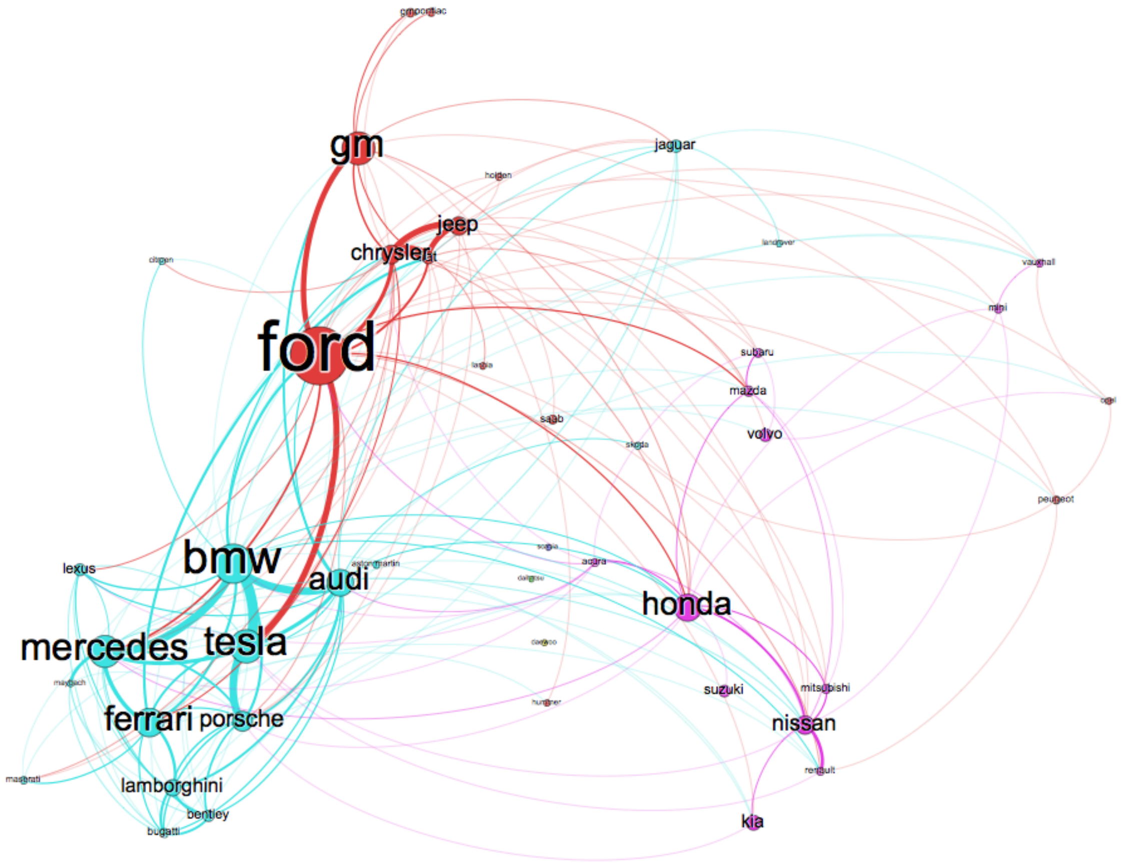

Chart 9: Automotive Industry Brand Co-Mention network

During this study experiments with different angles on brand co-mentions were conducted. One of the experiments was to look at single industries. Chart 9 above shows the car industry Brand Co-Mention network as an example. It reveals a luxury car club in light blue, a made-in-america-club in red and an asian-car-club in purple. It also shows that some brands are barely connected and potentially lacking a proper brand positioning. Further insights could be yielded by looking at co-mentions of car types which was not part of the study. This would reveal potential shortcomings of brands in certain car segments as well as what car models compete with each other. As mentioned before the underlying data is not limited to any geography however conducting this analysis for a certain geographic market would show a different picture and would be more insightful for management.

Another experiment conducted during this study was to capture brand mentions including personal and ethical values in tweets e.g. Safety, Luxury, Power, Quality and so on. The intention was to get an understanding of how a brand is perceived by consumers. The results didn't uncover pioneering insights. However adjustments in the methodology and process could improve the results and might reveal more valuable insights. A potential application for this is to look at changes over time and use the data as an instrument for measuring brand effectiveness and competitive intelligence. An image of the test results created with force atlas 2 can be found in the attachments.

By changing the scope of co-mentions from tweet to timeline the results draw a picture of what set of brands users are talking about in general. Although this data was extracted in this project, the first results were not covered in this report as different tools and approaches might be needed to extract meaningful insights. However the first observation was that brands have a much higher connectivity and a lot of connections outside of their industry. Which makes sense as users can tweet on different days about different brands. An interesting subsequent research question would be to understand what kind of users are tweeting about what kind of sets of brands. This question leads me to my actual research interest, behavioural segmentation.

There are plenty of existing segmentation models which are twitter specific or more general marketing models. However one area that I'm particularly interested in is to study the impact of emotions and personal values on buying decisions and brand preferences. One segmentation model which was created on the basis of neuroscience knowledge is the limbic map[14]. The basic concept behind it is to segment users by motivations. By understanding the motivation why users are interested in products or content on top of what they are interested in would allow better predictions for recommender systems, advertising targeting and more. I proved in my undergraduate thesis that this kind of model works well when studying and predicting the adoption of innovative products. However for most existing neuromarketing models the data collection is done via an offline survey panel. A potential MSc project could be to find a way to infer motivations and personal values from content that users create and consume on social media and/or on the web.

Finale Notes

During this study it was shown that looking at brand co-mentions on twitter can be very insightful and reveal valuable data points for brand management. Even though the data only reflects 3 days during Christmas of 2014, the results indicate an authentic picture for many industries. Nonetheless to make the data more reliable and statistically significant it would be

[14] Limbic Map by Gruppe Nymphenburg - http://www.nymphenburg.de/292.html

necessary to capture data over a longer period of time and ideally limit it to one geographical region.

References

Datasources

- http://en.wikipedia.org/wiki/Category:Lists_of_brands
- http://fanpagelist.com/
- http://help.sentiment140.com/for-students/
- http://snap.stanford.edu/data/twitter7.html

API’s, Libraries and Tools

- https://dev.twitter.com/rest/public
- https://dev.twitter.com/rest/reference/get/search/tweets
- https://dev.twitter.com/streaming/overview
- https://github.com/tweepy/tweepy
- https://networkx.github.io/
- http://gephi.github.io/
- http://www.r-project.org/

Appendix

Project File Structure

Path	Description
./	Main Exec Code
./twitter	Twitter API Handlers
./input_data	brand_keywords.json >> Brand keywords motivation_keywords.json >> personal values keywords
./input_data/brand_search_results	results of the brand queries / one file per brand query
./input_data/user_train	final data that was used to create the edges between brands / one file per user
./results/	results from extracting the edges and nodes of the tweets
./results/gephi/	gephi graph file and network property calculations
./results/rich club/	rich club result from NetworkX python library
./results/r/	R script and files

Top 20 Nodes

Label	Mentions	Degree	Closeness	Betweenness	Eigenvector
youtube	33.118	296	1.44	15,403	100%
twitter	22.205	254	1.52	9,321	92%
ebay	26.395	247	1.55	13,364	83%
google	24.121	237	1.57	7,445	90%
apple	23.410	219	1.60	5,364	86%
amazon	15.538	218	1.61	6,890	83%
iphone	23.046	203	1.64	3,506	83%
facebook	16.017	197	1.65	4,447	79%
android	18.923	193	1.66	3,182	81%
microsoft	11.317	178	1.70	2,240	79%
sony	17.326	171	1.71	3,249	73%
ipad	9.461	157	1.75	1,841	69%
windows	7.350	148	1.77	1,757	70%
instagram	7.201	136	1.78	3,804	56%
sky	2.982	127	1.79	3,263	55%
samsung	9.814	126	1.81	960	62%
yahoo	2.541	122	1.82	1,289	58%
reuters	5.860	121	1.81	1,402	54%
forbes	3.538	118	1.83	1,125	57%
hp	3.553	115	1.85	1,191	54%