# Data Driven Social Analytics
# Final Project: The Vent Dataset

Julián Lopez Baasch

## A) Introduction:

Vent is a semi-anonymous social networking that lets users share their feelings and frustrations without the fear of a negative backlash. It encourages users (also referred to as venters) to "express and share their feelings with people who care", without the worry of being insulted, de-friended or upsetting people they know. Apart from posting, users are exposed to the feeds of other users' vents, and they interact with them by reacting to their vents via a set of emotion.

The Dataset comprises more than 33 millions of vents posted by 934.095 users together with their social connections. Each vent has an associated emotion. There are 705 different emotions, organized into 63 "emotional categories". Moreover, the Dataset includes the social graph of Vent, containing the directed friendship links between users.

**Research Questions:**

The purpose of this study is to attempt to answer three main questions:

- RQ 1: Are Vent users proned to connect with users with similar emotional profile? Previous studies in a variety of social media contexts have shown evidence that users tend to interact and associate with others expressing similar emotions, referred as the phenomenon of "emotional homophily".
- RQ 2: Is there evidence of emotion contagion through users of the social graph of Vent? In this project we try to answer how emotions can be diffused across the links connecting individuals. Although the phenomenon of emotional contagion is well study in laboratory experiments, it is unexplored in the context of social networks.
- RQ 3: What can we learn about the affective mechanisms and relations between emotions? In order to answer this question we analyze the similarity and co-occurrence between different emotions making advantage of the temporal information provided in the Dataset.

For this purpose, we should first analyze and discuss the strong and weak points of the Dataset. This will be helpful in order to conceive the capabilities and limitations of the Dataset in terms of answering the stated Research Questions.

**Weak Points:**

- The Dataset does not provide temporal information about the social graph of Vent. In other words, the underlying network is static. Some notion about the link formation through time would be useful to study the relationship between the network structure and the Vent feed and vice-versa.
- There is no data about the interaction between users in the social network. Although Vent users can interact with other users by commenting or reacting via a set of emotion specific reactions, the Dataset only includes the amount of reactions for each vent. Therefore, the reactions available are anonymous, and lack of content.
- Users may use the platform for different reasons. While some users may find in Vent a place to expose their emotions, others may use it for promoting personal interests, like publicity or imposing political agenda. We do not know if this is true because we do not have the annotated text associated to the data.

**Strong Points:**

- The Dataset can provide a baseline corpus for emotion analysis of user-generated text. It is the largest annotated Dataset of text with effects.
- The most important feature of the Dataset is the existence of self-reported "ground truth" affect annotation associated with each text. In this sense, the labelling has been provided by the authors of the texts.

# B) Data Exploration:

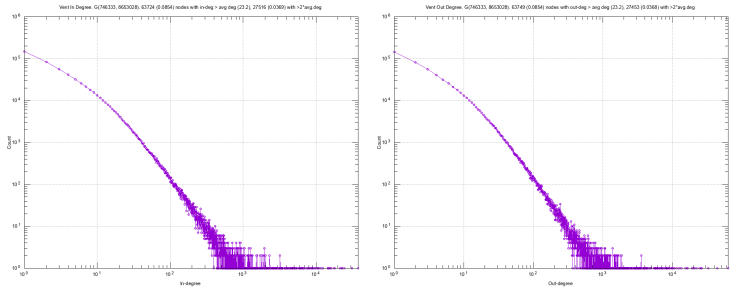
We now present an exploratory analysis of the Vent dataset. We first study the social graph of the Vent dataset, and we do a social network analysis (SNA) upon it. We compute the most common metrics of SNA in order to have a notion about the network we are dealing with. We then briefly analyze the users activity and the usage of the different emotions at the level of vents, and we propose a useful filtering of the dataset based on the results after the exploration. Finally, we attempt to answer the thre stated RQs.

# 1 Network Analysis and Data Preparation:

| Nodes | Edges | $< k >$ | D | Diam | Eff. Diam | k-core | C. Coeff. |
|---|---|---|---|---|---|---|---|
| 746333 | 8653028 | 21.2747 | 2.26x10$^{-}$5 | 11 | 3.891798 | 123 | 0.153645 |

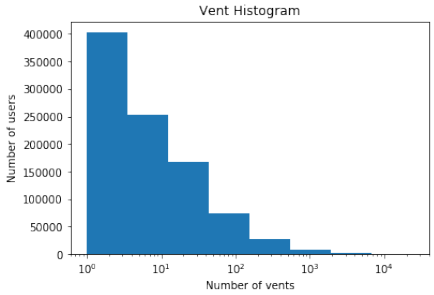
The table above shows the SNA metrics for the Giant Component (GCC) extracted from the social graph of Vent. The GCC has 746333 nodes, 8653028 edges and and average degree ($< k >$) of 21.27. It is noteworthy that the maximum k-core raises to 123, which means that the GCC is deeply connected. Also, we notice that the diameter, which is of 11, is much bigger than the 90% effective diameter (3.89), which suggests that the nodes in the GCC might present preference to connect with nodes with high degrees, a phenomenon we will cover later. Also, this result might be related to the fact that the GCC has a high maximum k-core.

The figure below shows the in-degree and out-degree distributions of the GCC. We observe the typical heavy tail behaviour in both graphs, with a vast majority of nodes linked to a small number of other users, and a small minority having more than 10K connections.



Considering the results exposed so far, we decided to filter out nodes with degrees less than 10 and nodes with degrees higher than 10K. This way, the resulting dataset will consist of users characterized by a 'normal' profile. Moreover, we consider that users that have more than 10K friends are likely to be bots, or employees at Vent.

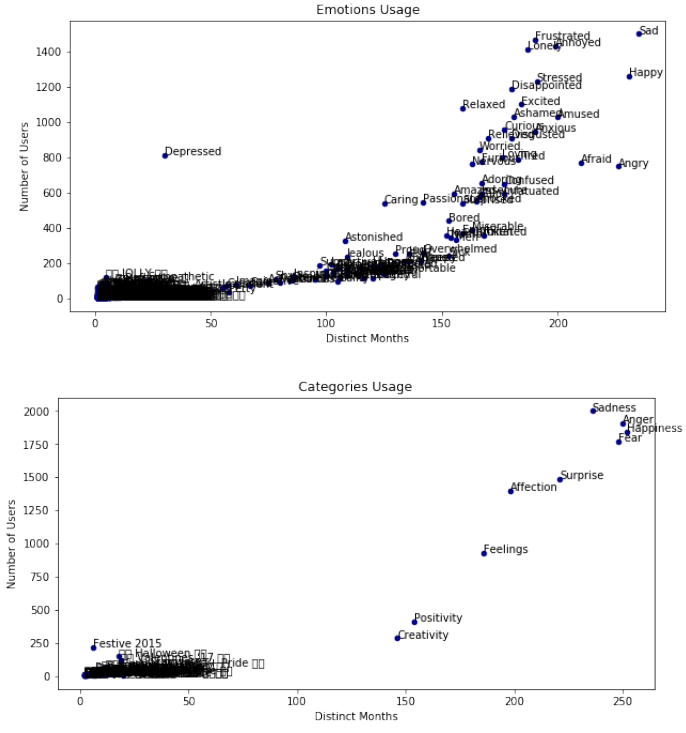
We now examine the user behaviour in the Vent platform. The figure below shows the histogram in Log-scale of vents per user. Again, we observe a heavy-tailed distribution, indicating that most of the users posted less than 10 vents while a small group of users posted more than 100.




Additionally, for the sake of robustness in the data, we filter out users with the lowest activity. Formally, we dismiss all users that presents less vents than average (36). This way, we can work with a consistent dataset without outliers in both the network structure and the user's activity.

We have described so far the data preparation and filtering we consider necessary in order to make a consistent analysis. Now we proceed to examine the emotional landscape offered by Vent platform to their users.

To determine the extent to which venters use this wide spectrum of emotions, we plot the relation between the availability of the emotions, and the users that choose this emotions for their vents. The figure below shows a scatter plot for the number of users and number of distinct months (235 in total) each emotions was available in the Vent platform.




We notice that while most of emotions concentrates near the origin of the graph, some of them are displayed separated and far from the origin. We consider the emotions last mentioned to be ordinary emotions people use to talk about (Happy, Sad, Afraid, Angry, Frustrated, Annoyed, Lonely, Disappointed, etc.). On the other hand, the remaining group of emotions near the origin correspond to be less frequent, as most of them are associated to a seasonal or festive day (Halloween, World Cup, LGTB Pride, etc), and we therefore treat them as outliers. Same can be said about the emotion categories, as this pattern even holds with more clarity. Only 9 of the 63 categories form part of the 'common' cluster: Sadness, Anger, Happiness, Fear, Surprise, Affection, Feellings, Positivity and Creativity. For some of the following analysis in this project, we make use of advantage of this results, as we will only consider a subset of emotions for treatment.

## 2 Assortativity:

Assortativity or homophily refers to the preference for the network's nodes to attach others that are similar in some way. In other words, similar nodes may be more likely to attach to each other than dissimilar ones. Though the measure of similarity may vary, we examine the assortativity in Vent's social graph in terms of degrees, emotions and number of reactions. We think that users not only may link to users who shares the same emotional profile, but that also most of users tend to connect with other users with similar degree values or popularity (number of reactions).

| **Degree** | **Attr.** | **Numeric** |
|---|---|---|
| -0.121 | 0.029 | -0.162 |

The table shows the three coefficients computed according to Newman (2003). For the attribute assortativity, the formula implemented is given by:

$$\frac{trace(M) - sum(M)}{1 - sum(M)}$$

where M is the joint probability distribution (mixing matrix) of the specified attribute (see NetworkX documentation). For the case of emotion assortativity, we defined M as the most frequent emotion vented by each user. Although this approach is arguably biased, we consider that is fair enough to associate the most frequent emotion for each user to their emotional profile. In the case of the reactions assortativity, we used the sum of reactions as a numeric attribute.

We observe that the degree assortativity is slightly negative. Similarly, the reactions assortativity. However, our network presents a slightly homophily according to the emotions. We believe this result is not trivial: emotional assortativity may be rather small, but it implies a shift from negative to positive respect to the degree assortativity used as benchmark.

A further suggested work would be computing the assortativity for different behaviours and features of the nodes. Furthermore, the approach can be extended and enriched with the provided temporal information. For example, it would be interested to see if Vent users have preference to connect with users that shares the same emotions at any given point in time, by taking the last vent posted by each user. Then, like in an inverted approach, it would be possible to see the temporal evolution of the assortativity coefficient. One more interesting analysis would be to study the assortativity within specific periods, like the World Cup, or political elections. For this kind of events, Vent had enabled specific emotions, for example, "Supportive"+*Flag*, with *flag* being the flag of any country participating in the World Cup.

## 3 Emotion Spread:

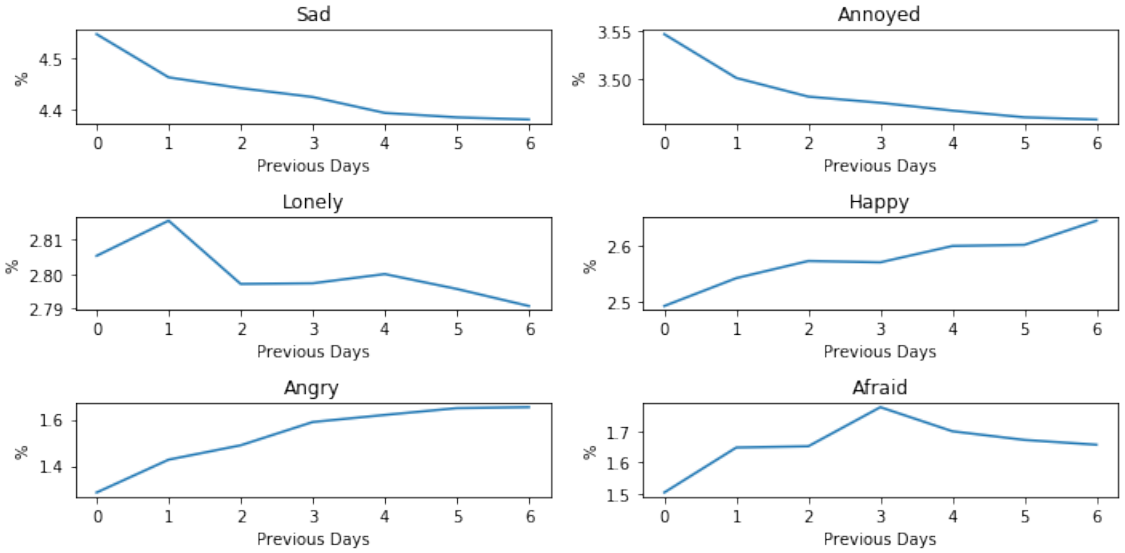
We now analyze the mechanism of emotion difussion. We propose an algorithm to estimate how a vent is influenced by the vents posted within a certain window by the neighbors of the owner of the vent. In other words, we want to estimate if vents emotions are correlated to the emotions of the previous vents in the neighborhood.

For this reason, we take a random vent and see the emotion associated with it. The next step is to look at the neighbors of the owner of the original vent, and look at their vents within a certain time window. We observe the emotions associated to the pool of this preceding vents, and calculate how many of them coincide with the emotion associated to the original vent. Finally, we calculate the ratio of coincidences in emotions and all the emotions in the pool of preceding vents. We formally describe below the algorithm implemented. We repeat this steps 5000 times and average in order to get robusts results.

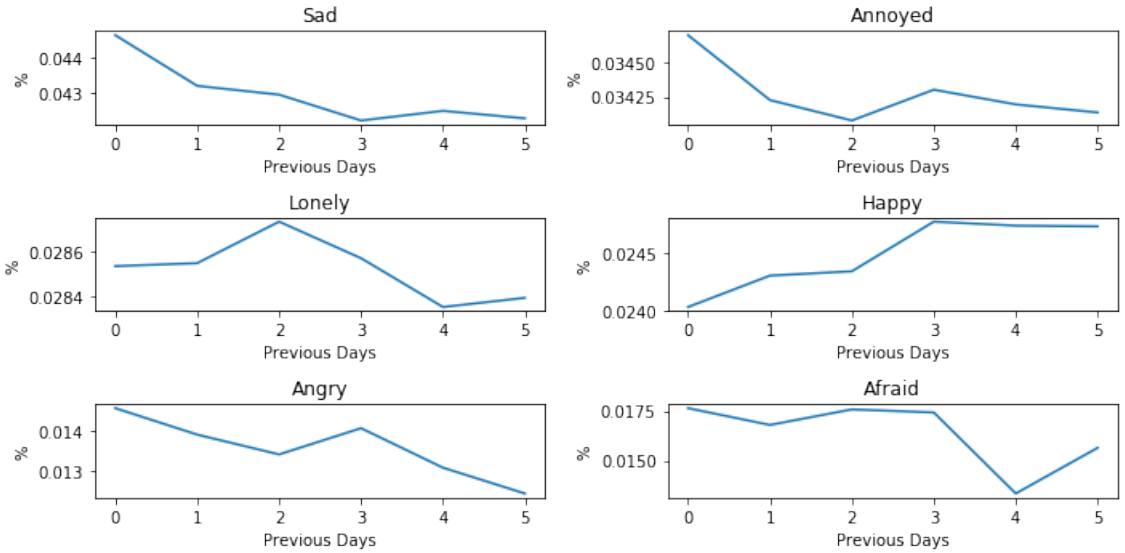
**Algorithm 1** Emotion Contagion

**for** *1 to n_simulaions* **do**
  vent $\leftarrow VENTS.Sample(1)$
  $time \leftarrow vent.Time$
  $emotion \leftarrow vent.Emotion$
  **for** *days = 1 to 7* **do**
    window $\leftarrow VENTS[(VENTS.Time < time)\text{and}(VENTS.Time >= time - days)]$
    $lenght \leftarrow len(window)$
    $ratio \leftarrow len(window[window.Emotion == emotion])/length$
    **end**
  **end**

We run the implementation for the 6 most common emotions: Sad, Annoyed, Lonely, Happy, Angry and Afraid. We used 7 time windows: from 1 day to 7 days. We run 5000 simulations. The results are shown below. We can observe that in the case of Sad and Annoyed, the preceding vents has an impact on the original vent. This result does not hold for the rest of the emotions.



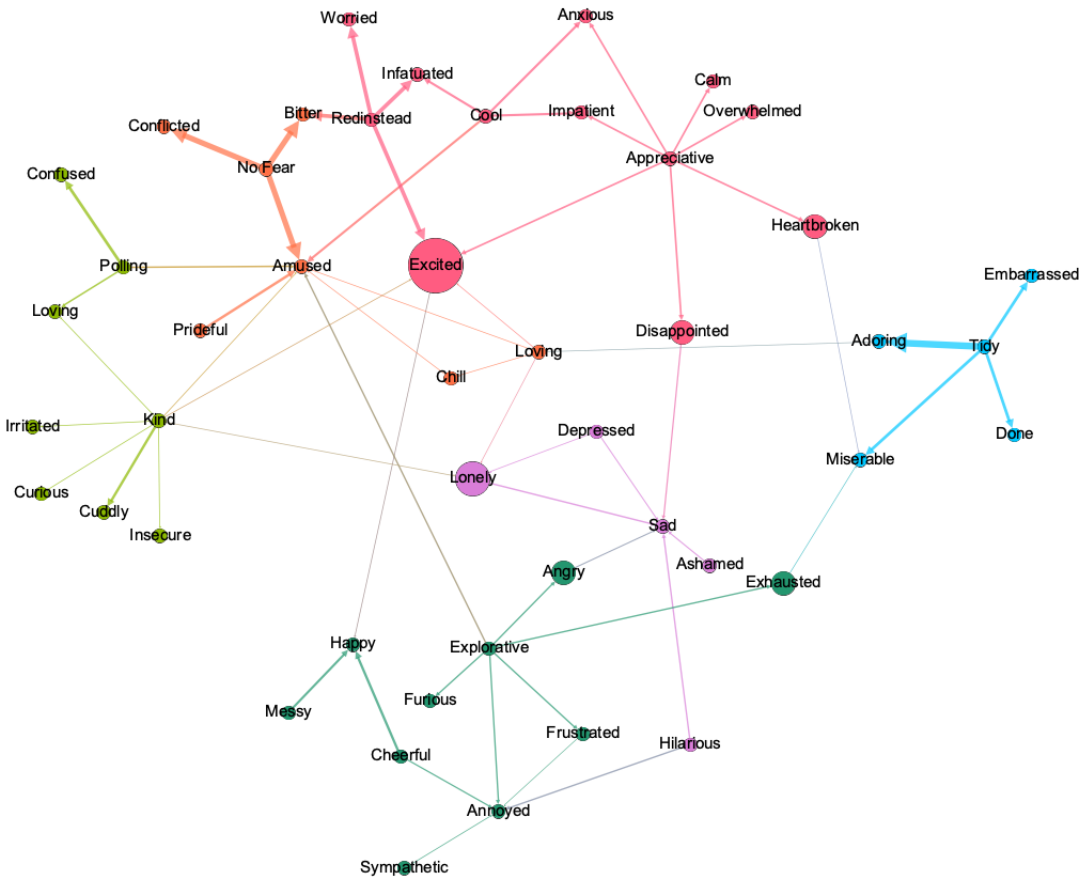

We then repeat the exercise but instead of using a a varying time window, a fixed, or rolling, window is implemented. In this case we use a window of 2 days, iterating over the previous 7 days of each vent sampled. The results are similar to the one shown before.



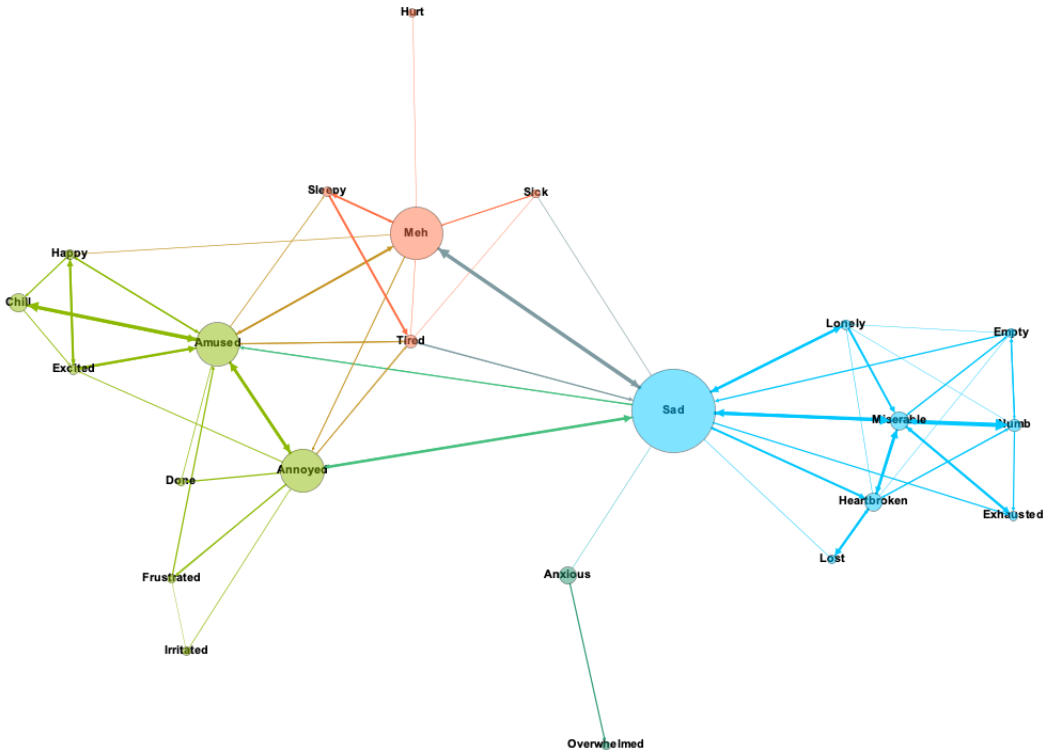
# 4 Emotion Transitions:

In this section we try to answer the last RQ. In order to study the affective mechanism and hidden transitions between emotions, a transition matrix is computed for a set of emotions. The approach used to get this matrix was the following: for each vent, we take the following vent by the same user. We then count the occurrences for each pair of transitions emotions. Finally we normalize the occurrences by the original emotions, resulting in a sort of transition probabilities.

We show in the following figure, the network resulting from applying the steps described above to the biggest ego-network. We filter out edges with a weight less than 0.08 for the sake of visualization. A lower threshold would imply in an almost fully connected network. Nodes are coloured according to the communities detected applying Louvain method, and the size of degrees correspond to the Betweeness Centrality. The network is dislpayed with a Yifan Hu layout. It is noteworthy that some communities are similar to the proper categorization of emotions, as most emotions are connected to emotions belonging to the same category. Also, it is interest to see that the most central nodes, like Lonely, Dissapointed or Exited, are less 'extreme' than emotions with lower centrality like 'Happy', 'Sad' or 'Miserable'




The figure below corresponds to a transition network obtained from a smaller ego-network, and using different threshold. We can see that the communities after applying Louvain method are considerably similar to the real categorization.




# C) Conclusion:

In this project, we studied the Vent Dataset and the interaction between the social network structure and the user's activities. Although the results obtained are not of the quality desired, we beleive that the Dataset has the potential to yield more insights about interaction between humans throuhgh an annotated emotional landscape. We have seen that the exposure to some emotions may influence users to post vents with same emotions. Although the analysis was taken upon random sample of vents, we suggest that within certain conditions, and in specific time periods, like political elections, the accumulation of emotions of certain type may catalyze the production of emotions of the same type. That kind of phenomenon occurs often in other social networks, and there is no reason to assume that in Vent things will be different.

In addition, we can say the Dataset provide a resource to study if users are prone to connect with other users with similar emotional profile. We computed the assortativity coeff upon all the network, and saw that the users are slightly assortative. A more detailed analysis and a stronger filtering would probably derive in higher homophily.

Finally, we suggest that the hidden transition mechanism behind emotions can be an useful tool to model NLP and emotion analysis. We did a first attempt to compute a transition matrix, getting little but modest results. Nonetheless, there is no doubt a more exhaustive study would yield to more complex models applicable to NLP.