Misery Loves Company: Happiness and Communication in the City
Supporting Information

Aamena Alshamsi Edmond Awad Maryam Almehrezi
Vahan Babushkin Pai-Ju Chang Zakariyah Shoroye
Attila-Péter Tóth
Iyad Rahwan
Masdar Institute of Science and Technology, Abu Dhabi, UAE
Massachusetts Institute of Technology, Cambridge, USA

March 6, 2015

A Sensitivity analysis of the homophily in the communication patterns of urban areas

We chose different percentiles to label areas as happy or unhappy. Then, we used two-way Analysis of Variance (ANOVA) to compare the effect of using the different percentiles on the strength of communication.

We tried percentiles 15, 20, 25, 40 and 50 and generated the results accordingly in tables 1, 2, 3, 4 and 5 and the interaction plots are depicted in Figures 1 respectively. In summary, homophily exists for all the chosen percentiles from 15% to 50%.

Main Effects			
Variable		F(1,52418)	Pr(>F)
Source (Caller)		0.533	0.465
Receiver		0.935	0.334
Source * Receiver		72.733	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.039	0.100
Happy	Unhappy	0.034	0.076
Unhappy	Happy	0.034	0.075
Unhappy	Unhappy	0.042	0.107

Table 1: ANOVA results of normalized communication (threshold 15%)

Main Effects			
Variable		F(1,98656)	Pr(>F)
Source (Caller)		0.051	0.821
Receiver		1.169	0.280
Source * Receiver		91.492	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.037	0.101
Happy	Unhappy	0.033	0.079
Unhappy	Happy	0.033	0.079
Unhappy	Unhappy	0.039	0.103

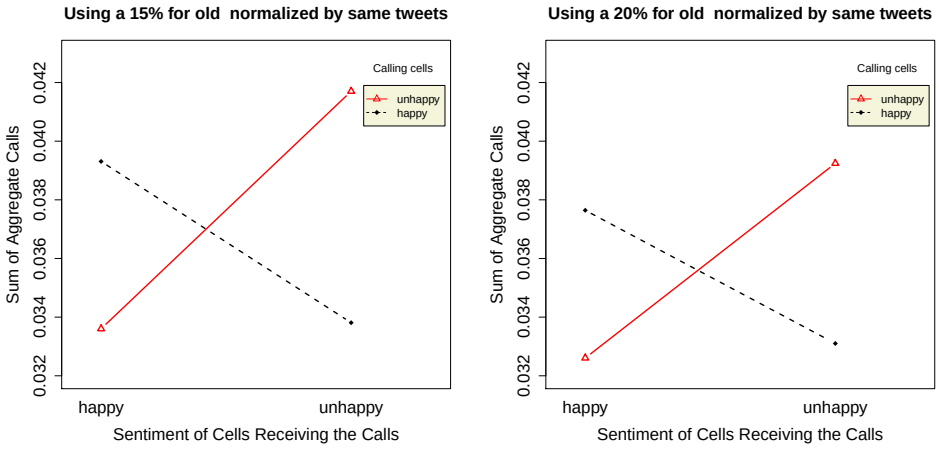
Table 2: ANOVA results of transformed and normalized communication (threshold 20%)

Main Effects			
Variable		F(1,163459)	Pr(>F)
Source (Caller)		4.403	0.036
Receiver		6.829	0.009
Source * Receiver		61.711	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.036	0.094
Happy	Unhappy	0.034	0.084
Unhappy	Happy	0.034	0.082
Unhappy	Unhappy	0.038	0.101

Table 3: ANOVA results of normalized communication (threshold 25%)

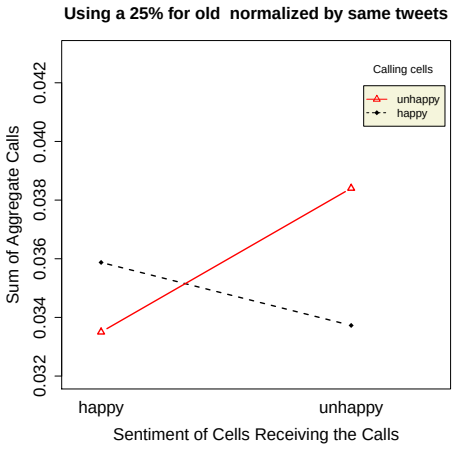
Main Effects			
Variable		F(1,452976)	Pr(>F)
Source (Caller)		8.222	0.004
Receiver		11.174	0.001
Source * Receiver		76.603	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.035	0.086
Happy	Unhappy	0.033	0.085
Unhappy	Happy	0.033	0.085
Unhappy	Unhappy	0.037	0.096

Table 4: ANOVA results of normalized communication (threshold 40%)

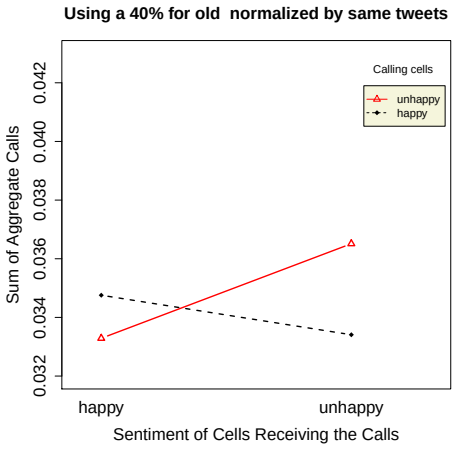


(a)

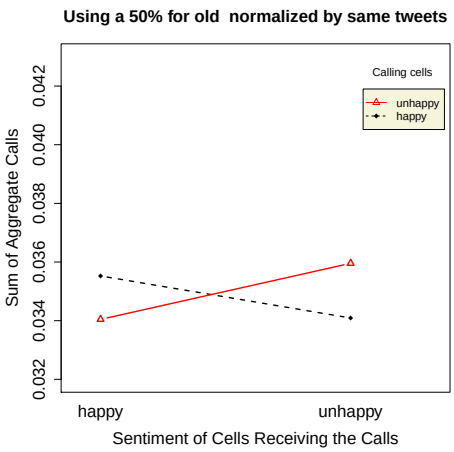
(b)



(c)



(d)



(e)

Figure 1: Interaction Plots of Data Using Different Percentiles: (a) 15% (b) 20% (c) 25% (d) 40% (e) 50%

Main Effects			
Variable		F(1,735801)	Pr(>F)
Source (Caller)		0.560	0.454
Receiver		0.961	0.327
Source * Receiver		62.703	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.036	0.089
Happy	Unhappy	0.034	0.089
Unhappy	Happy	0.034	0.088
Unhappy	Unhappy	0.036	0.096

Table 5: ANOVA results of normalized communication (threshold 50%)

B The Distribution of Communication Data

As we reported in the main paper, the distribution of the outgoing communication of areas (as shown in Figure 1(a) in the paper) is long tailed and could be well approximated by a power law. We examined the distribution of aggregate communication between areas and we found that it is also skewed by nature as shown in Figure 2(a). This might violate the normality assumption of ANOVA which requires the distribution of the residuals to be normal (check Figure 2(b)). Since the data is skewed by nature, we transformed the communication data using the natural logarithm. The transformation has reduced the skewness of the distribution of the communication and the residuals as shown in Figure 2(c&d).

After we transformed the data, we ran ANOVA to check whether our findings of homophily still hold. We found that homophily still holds for many thresholds from 10% to 50% as shown in Tables 6, 7, 8, 9 and 10. The interaction plots are shown in Figure 3.

C Diversity of happiness within each area

We examined the diversity of happiness within each area by calculating the standard deviation of happiness in each area, and upon plotting the distribution of the standard deviations in Figure 4, one can see that only few cells have standard deviations higher than 1.5. Hence, the existence of an area full of tweets with only scores 1 and 9 (i.e. having a standard deviation around 4) is very unlikely. In fact, the existence of areas with only scores less or equal to 3 and higher or equal to 7 (i.e. having a standard deviation of at least 2) is unlikely.

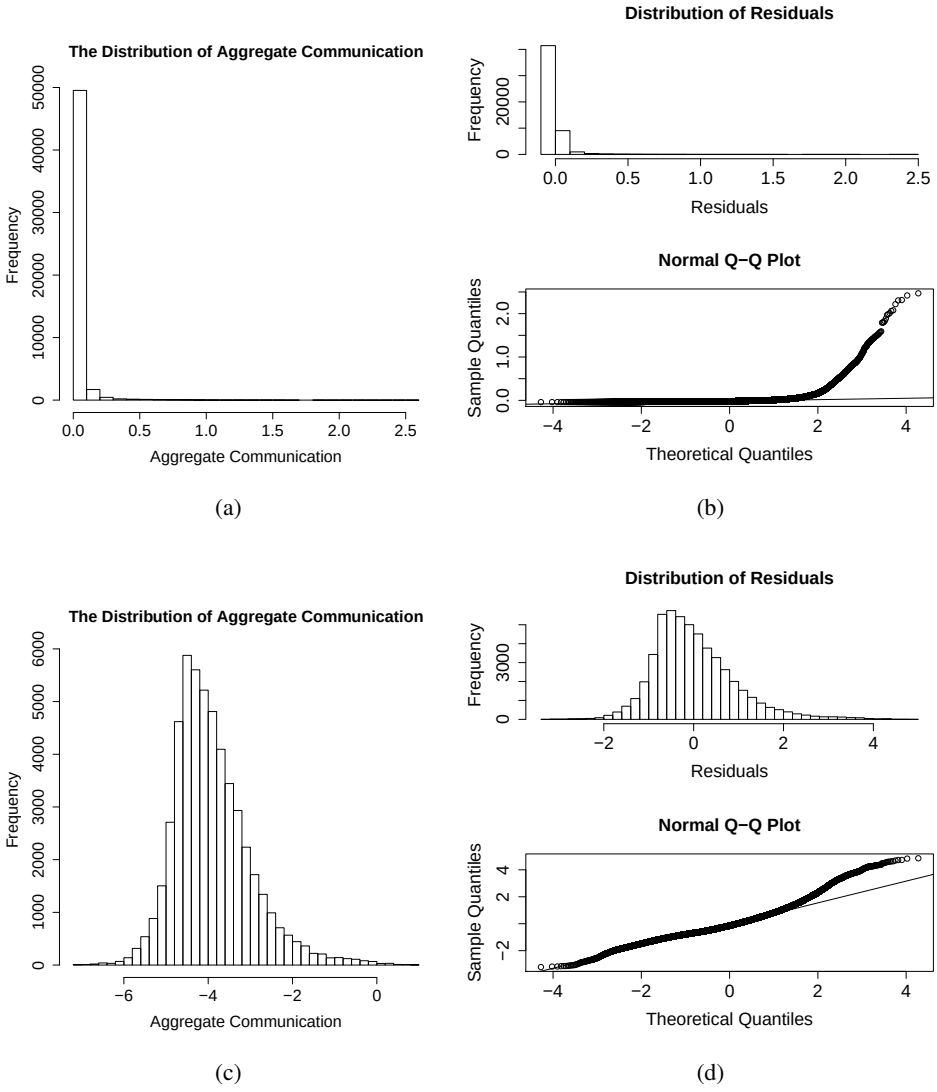


Figure 2: Normality Assumption of ANOVA: We use the data that include areas with lowest and highest 15% happiness scores (a) The distribution of aggregate communication between areas (b) The Q-Q plot of residuals after running ANOVA (c) The distribution of *log-transformed* aggregate communication between areas (d) The Q-Q plot of *log-transformed* residuals after running ANOVA

Main Effects			
Variable		F(1,52418)	Pr(>F)
Source (Caller)		0.022	0.882
Receiver		0.215	0.643
Source * Receiver		47.468	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.039	0.100
Happy	Unhappy	0.034	0.076
Unhappy	Happy	0.034	0.075
Unhappy	Unhappy	0.042	0.107

Table 6: ANOVA results of transformed and normalized communication (threshold 15%)

Main Effects			
Variable		F(1,98656)	Pr(>F)
Source (Caller)		0.118	0.731
Receiver		6.424	0.011
Source * Receiver		40.509	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.038	0.101
Happy	Unhappy	0.033	0.079
Unhappy	Happy	0.033	0.079
Unhappy	Unhappy	0.039	0.103

Table 7: ANOVA results of transformed and normalized communication (threshold 20%)

Main Effects			
Variable		F(1,163459)	Pr(>F)
Source (Caller)		14.570	0.000
Receiver		42.960	0.000
Source * Receiver		13.850	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.036	0.094
Happy	Unhappy	0.034	0.082
Unhappy	Happy	0.034	0.084
Unhappy	Unhappy	0.038	0.101

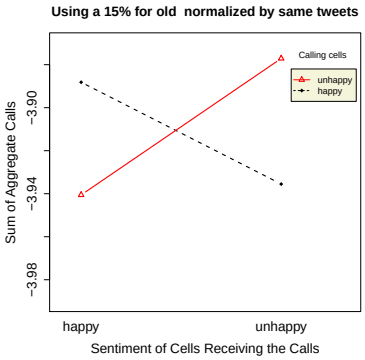
Table 8: ANOVA results of transformed and normalized communication (threshold 25%)

Main Effects			
Variable		F(1,452976)	Pr(>F)
Source (Caller)		1.333	0.248
Receiver		12.760	0.000
Source * Receiver		65.283	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.035	0.086
Happy	Unhappy	0.033	0.085
Unhappy	Happy	0.033	0.085
Unhappy	Unhappy	0.037	0.096

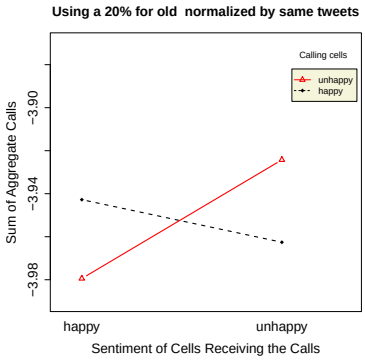
Table 9: ANOVA results of transformed and normalized communication (threshold 40%)

Main Effects			
Variable		F(1,735801)	Pr(>F)
Source (Caller)		12.421	0.000
Receiver		2.263	0.133
Source * Receiver		70.119	0.000
Pairwise Comparison			
From	To	Mean	Standard Deviation
Happy	Happy	0.036	0.089
Happy	Unhappy	0.034	0.089
Unhappy	Happy	0.034	0.088
Unhappy	Unhappy	0.036	0.096

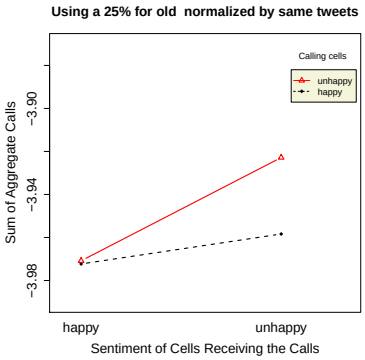
Table 10: ANOVA results of transformed and normalized communication (threshold 50%)



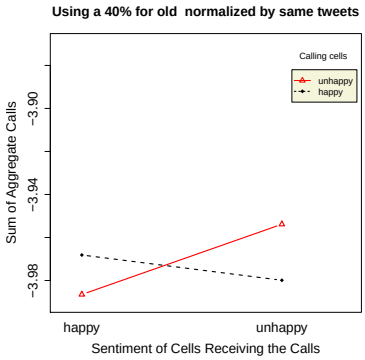
(a)



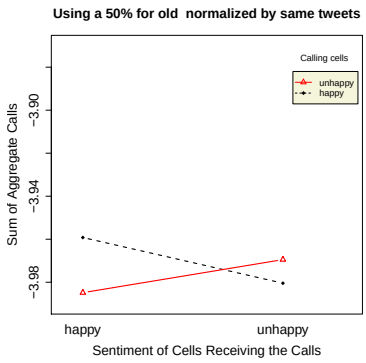
(b)



(c)



(d)



(e)

Figure 3: Interaction Plots of Transformed Data Using Different Percentiles: (a) 15% (b) 20% (c) 25% (d) 40% (e) 50%

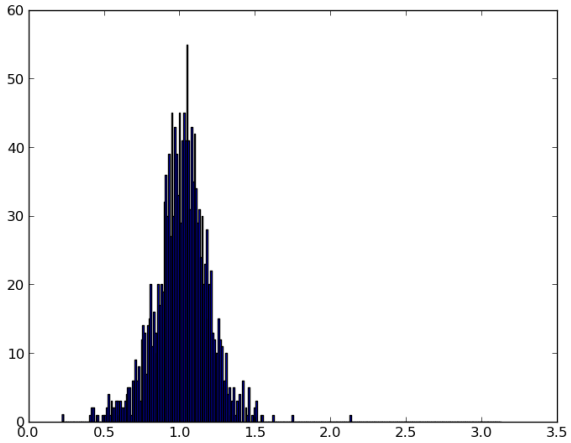


Figure 4: The standard deviation of happiness within each cell

D Weighted Assortativity Mixing

To quantify the level of assortativity mixing in the previously constructed network, we used a weighted version of the assortativity coefficient defined by [4]. We adjusted the way the coefficient is measured by incorporating the weights of the edges between nodes. Let $Y = \{happy, unhappy\}$ be the set of considered types, and let e_{ij} be the fraction of weights of edges that connect a node of type $i \in Y$ to another node of type $j \in Y$. Then:

$$\sum_{i,j} e_{ij} = 1$$

Let $a_i = \sum_j e_{ij}$ be the aggregated fractions of weights of edges that connect a node of type $i \in Y$ to all nodes, and $b_j = \sum_i e_{ij}$ be the aggregated fractions of weights of edges that connect all nodes to nodes of type $j \in Y$. Then, the weighted assortativity coefficient is calculated as follows:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

A value of $r = 0$, indicates that there is no weighted assortative mixing ($e_{ij} = a_i b_j$), while a value of $r = 1$, indicates that there is a perfect weighted assortative mixing ($\sum_i e_{ii} = 1$).

Note that while all of our analysis in the paper excludes self-edges, the assortativity function assumes the existence of self-edges (it normalizes w.r.t all edges including self-edges). Thus, we show the assortativity results for the cases when self-edges are included and when they are excluded (for percentages 15, 20, 25, 40, and 50). In the paper, we report the value for 15% with self-edges. Table 11 shows the results.

Percentile	Without Self-edges	With Self-edges
15%	0.099	0.216
20%	0.075	0.165
25%	0.048	0.126
40%	0.033	0.081
50%	0.024	0.062

Table 11: Weighted assortativity coefficient for different percentiles and considering with(out) self-edges.

E Correlations between Happiness, Communications Activities and Centrality Measures

Figure 5 shows the distribution of incoming/ outgoing calls/SMSs and the Internet traffic per area. Each of the five distributions is long tailed, and part of it is well approximated by power law. Additionally, one can observe the exponential cutoff at the tail of all distributions, which is likely attributed to constraints on time, attention, bandwidth, etc.

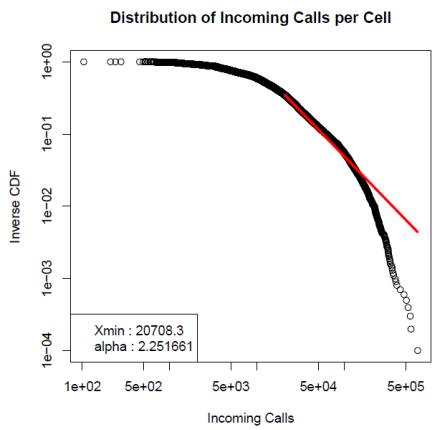
Figure 6 compares the happiness score of cells with the amounts of incoming and outgoing calls and SMSs and Internet traffic of these cells. Although, as expected, the various communication metrics are highly correlated, none of them is a significant predictor of happiness.

Figure 7 compares the happiness score of cells with the centrality measures (e.g. in-Degree, betweenness, etc.) of calls' network. Again, as expected, the various centrality measures are highly correlated. However, they fail to exhibit significant correlation with happiness. This implies that mere ability to mediate information flow is not sufficient to influence happiness.

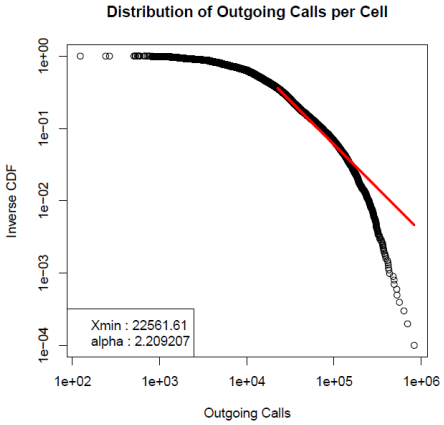
Figure 8 shows the network of calls (links) between the different areas (nodes) in Milan. Red nodes represent happy areas and blue nodes represent unhappy areas. Visual inspection highlights the presence of communities of areas that are dominated by a particular class (happy or unhappy).

F Correlations between Happiness and Geographic Distance

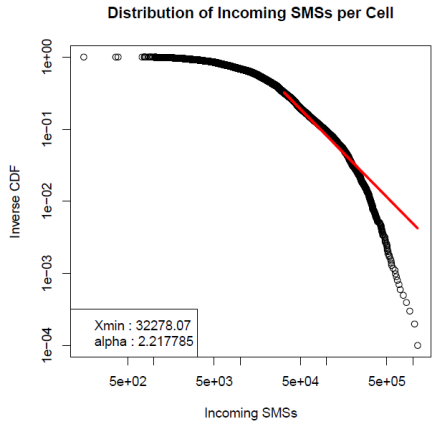
We studied the effect of geographic distance on cell happiness score, by showing the correlation between cell own happiness score and average neighbor's happiness score as a function of the distance between them. We chose to use Spearman's rank correlation since it has less constraints (e.g. the input data can be non-interval scale) and, more importantly, it is more tolerant to the outliers compared to Pearson correlation. If the correlation coefficient of a cell's own happiness score and neighbors' happiness score show no relationship, we can conclude that homophily found



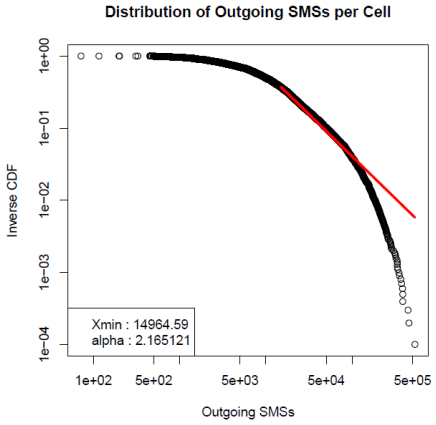
(a)



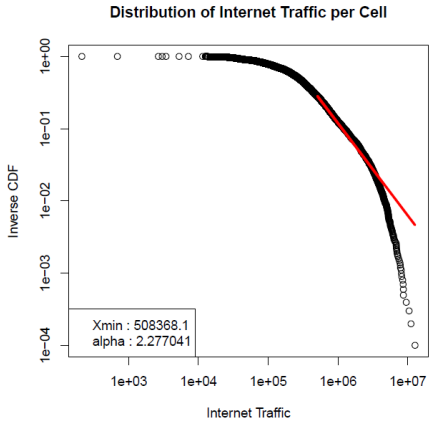
(b)



(c)



(d)



(e)

Figure 5: **(a) Distribution of incoming calls per area**. **(b) Distribution of outgoing calls per area**. **(c) Distribution of incoming SMSs per area**. **(d) Distribution of outgoing SMSs per area**. **(e) Distribution of Internet traffic per area.** All the five distributions are long tailed. An exponential drop at the tail of the distribution can be noticed as well, and is likely attributed to some constraints.

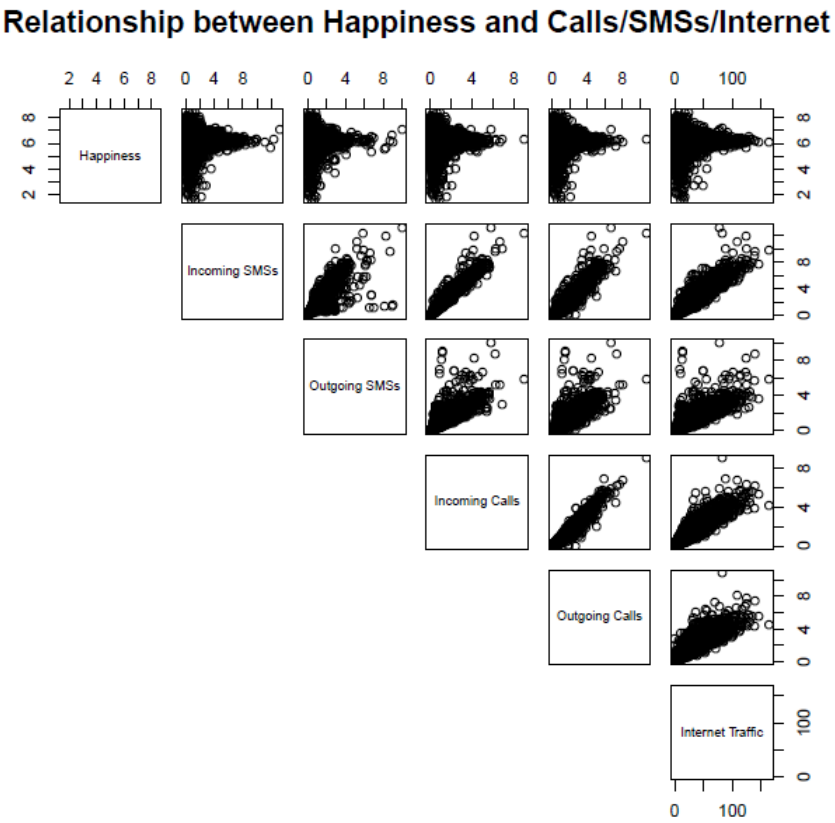


Figure 6: **Scatter plot matrix of correlations between happiness and calls/SMSs/Internet.** Points represent cells with non-zero Happiness score. Communication metrics are highly correlated with each others. However, none of them seems to make a good predictor of happiness.

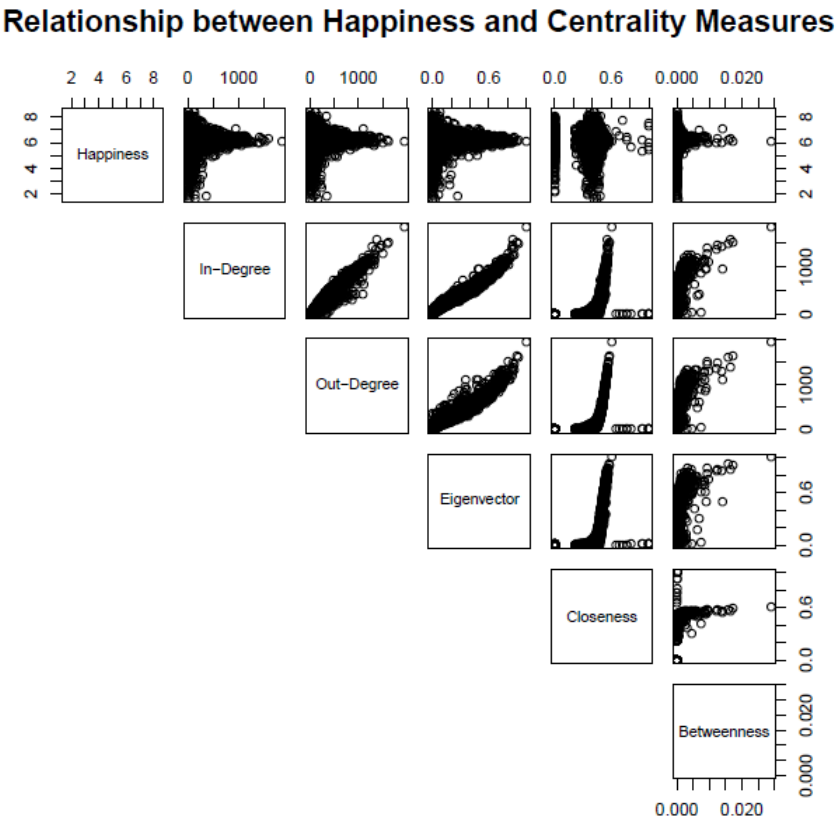


Figure 7: **Scatter plot matrix of correlations between happiness and centrality measures of calls' network.** Points represent cells with non-zero Happiness score. None of the centrality measures shows significant correlation with happiness.

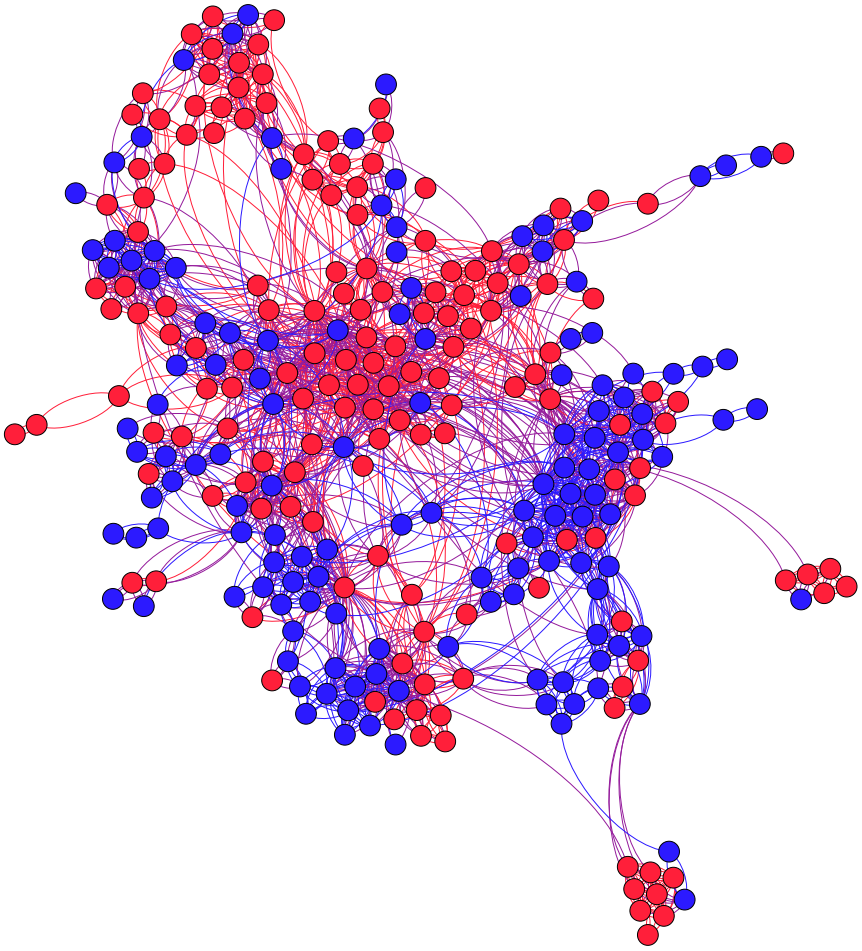


Figure 8: **Network of calls between happy/unhappy areas in Milan.** Nodes represent areas, and directed edges (i.e. arcs) represent calls between these areas. Happy areas are in red, while unhappy areas are in blue. A red arc connects two red nodes, a blue arc connects two blue nodes, and a purple arc connects a red node to a blue one, or vice versa.

in our results do not depend on the geographic distance between cells' location.

Figure 9 shows the correlation values given Euclidean (a),(c) and Manhattan (b),(d) metrics for distances calculations. Sub-figures (c) and (d) show the cases where only the cells with the highest and lowest 15% happiness values are kept. In all cases, one can notice that the correlation between the happiness of a cell and its neighbors exhibits frequent fluctuations around zero on the distance scale. This suggests that homophily based on proximity is very insignificant.

G Studying Homophily on Community Level

We applied a modularity-maximizing algorithm to categorize cells into communities. In our experiment, multi-level modularity optimization algorithm [1] is chosen to determine communities based on the telecommunication data. Modularity is a quality measure for graph clustering proposed by Newman [6, 5]. For a graph $G = (V,E)$, given a community structure, and a collection of disjoint subset of vertices $V = \{V_1, V_2, V_3 \ldots, V_l\}$, the modularity of community structure is defined as:

$$Modularity(V) = \frac{1}{2w} \sum_{ij} (W_{ij} - \frac{w_i w_j}{2w}) \delta_{ij}$$

where

$$\delta_{ij} = \begin{cases} 1 & (\text{if } i, j \text{ are in the same community}) \\ 0 & (\text{otherwise}) \end{cases}$$

$$\begin{aligned} i, j &\in V \\ w_i &= \text{weighted degree of node } i \\ w_j &= \text{weighted degree of node } j \\ W_{ij} &= \text{weight of link from node } i \text{ to node } j \\ w &= \frac{1}{2} \sum_{ij} W_{ij} \end{aligned}$$

As we can see, modularity is defined as the fraction of links that fall within communities minus the expected value of the same quantity if links are assigned at random, conditional on the given community memberships and the degrees of vertices. The value of modularity is between -0.5 to 1. When the modularity is high, it shows that the community structure have strong intra-community interaction and weak inter-community interaction. On the contrary, the community structure has low modularity when intra-community interaction is weak and the inter-community interaction is strong.

Usually, we assume cells in the same community have strong interactions. Therefore, we should apply some algorithm to maximize modularity for finding

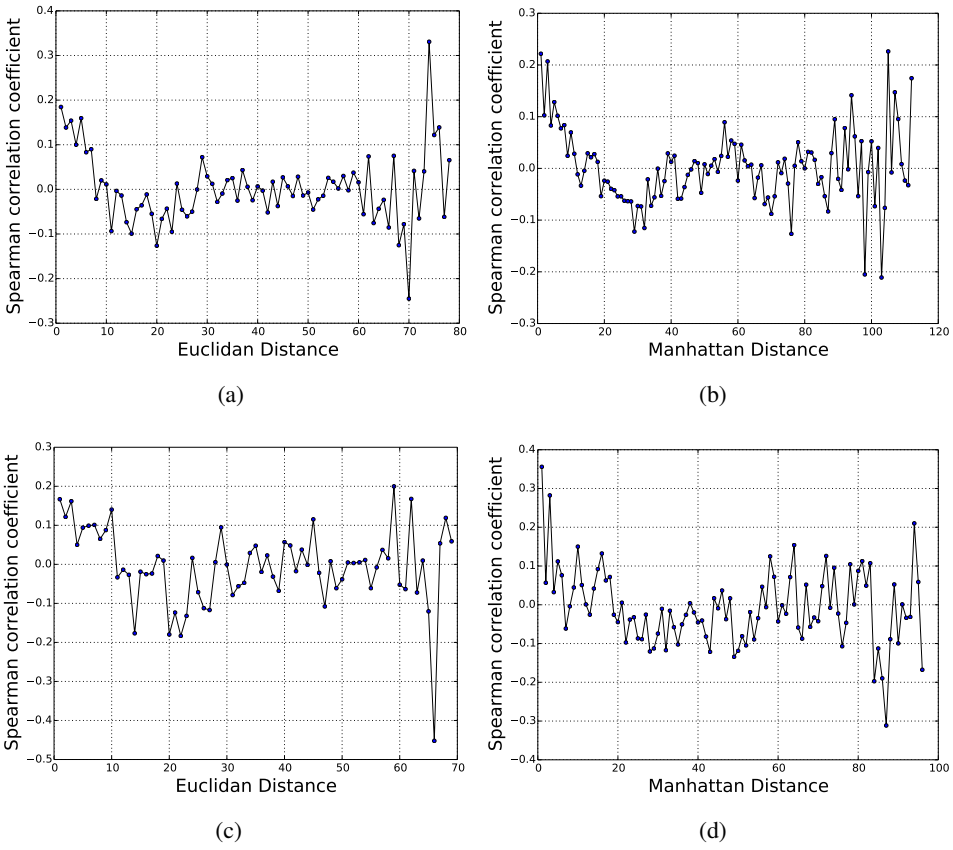


Figure 9: **The Spearman correlation coefficient between cell's happiness and its neighbor with different distance: (a) Euclidean distance, no threshold**, **(b) Manhattan distance, no threshold**, **(c) Euclidean distance, threshold = 15%**, **(d) Manhattan distance, threshold = 15%**. The correlation between the happiness of a cell and its neighbors exhibits frequent fluctuations around zero on the distance scale, which suggests the insignficant effect of proximity on homophily.

the best community structure. The multi-level modularity optimization, also called Louvain method, is a greedy optimization method that attempts to optimize the modularity of the network. The procedure can be divided into two iterative phases. First, each node will be assigned to a distinct community. As a result, the initial community number will equal to the number of nodes the graph contained. Then, for each node k, we evaluate the gain of modularity by removing node k from its community then placing it in a neighbor's community. The node k will be placed in the neighbor's community if this increases modularity. Otherwise, the node k will remain in the original community. If no more merging procedure can be done to increase the modularity, then the current communities' structure will be the best one to represent network community. Previous research has shown that multi-level modularity optimization is simple, efficient and easy-to-implement method for identifying communities in large networks [3, 2, 7]. The time capacity is $O(n \log n)$ and the runtime is near linear when the number of nodes is approximately equal to number of edges (sparse network).

After we found communities, we studied the effect of community size on the average and the standard deviation of a community's happiness score. If the standard deviation is small, then there is evidence for homophily on community level. Additionally, we are interested in finding whether average cell happiness score will change as the size of the community changes. The community size is defined as the number of cells in it. We consider different percentiles for labeling cells as happy/unhappy.

For comparison, we generate random communities of similar sizes of the communities we have. These are formed by randomly assigning cells into communities. We run the process for 200 times then take the average value then label both real and random communities. The results are shown in Figure 10 and Figure 11.

Figure 10 shows that small-size communities seem to have slightly higher happiness score on average compared to the random communities. Overall, the happiness value tend to decrease with the increase in community's size, for all percentages.

Figure 11 shows that detected communities seem to have lower standard deviations than random communities. This provides an evidence for the existence of homophily on community level. With the increase of the community size, the standard deviation of happiness score increases for both detected and random communities. As the blue lines in Figure 11 (a)-(b) show, the standard deviation slightly increases for 15% and 20% as the size of the community increases, while the blue lines show that the standard deviation stays constant or decreases for the other percentages (Figure 11 (c)-(e)) as the size of the community increases.

H Number of Tweets, Cells, and Communication Links

Table 12 shows the number of tweets, cells and links during each stage of preprocessing.

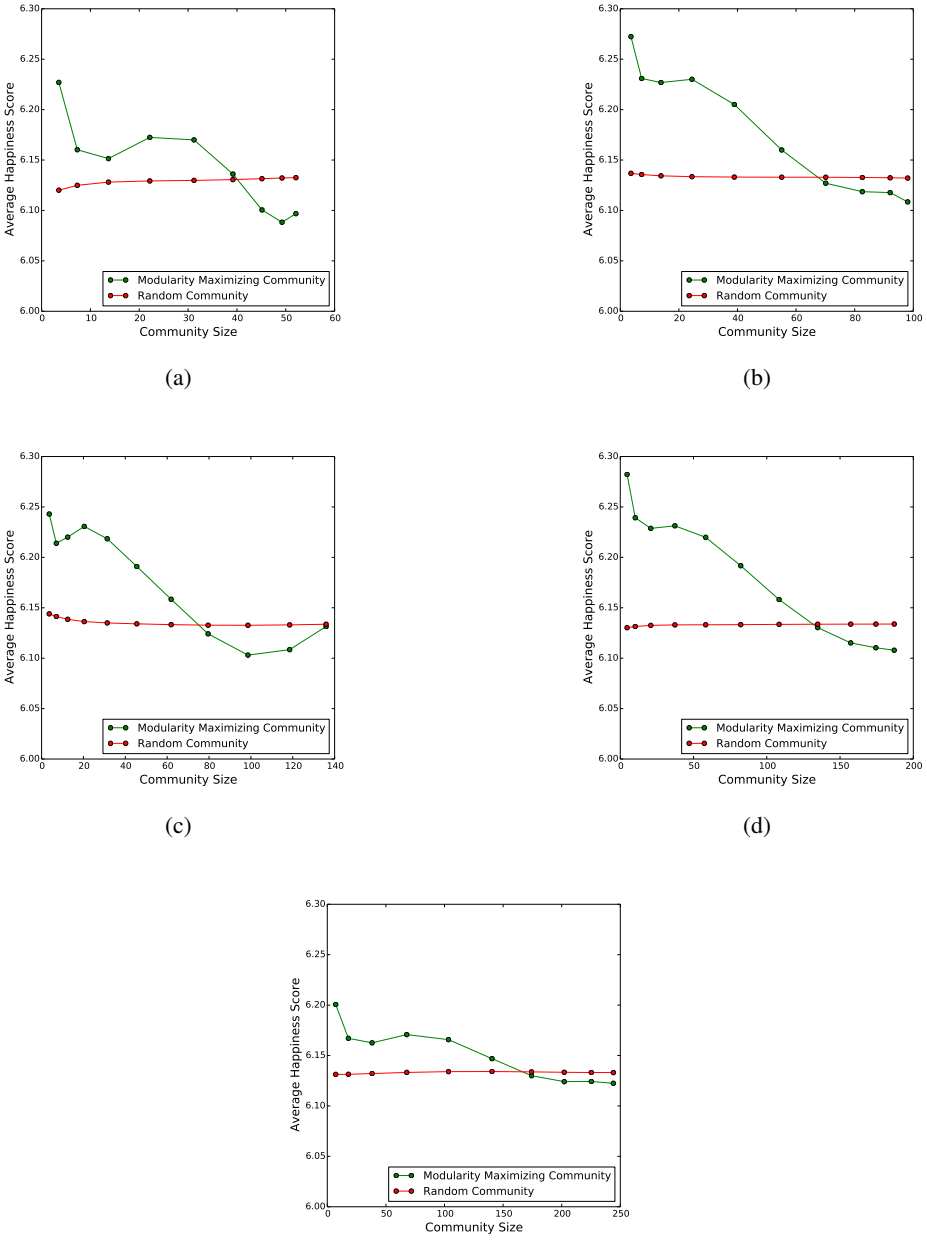
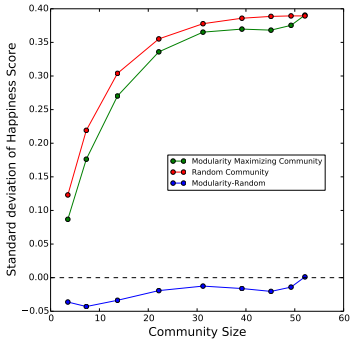
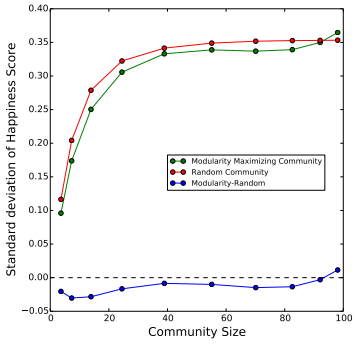


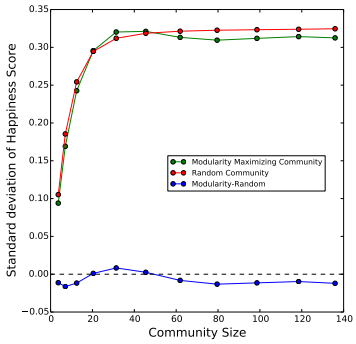
Figure 10: The average happiness of a community as a function of its size. **(a) threshold =15%**, **(b) threshold =20%**, **(c) threshold =25%**, **(d) threshold =40%**, **(e) threshold =50%**. Each point represents a single community and the y axis denotes average happiness score of all cells within this community. The green points represent communities detected by the modularity maximizing algorithm, while the red points represent the random communities.



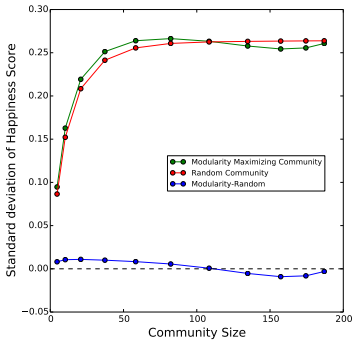
(a)



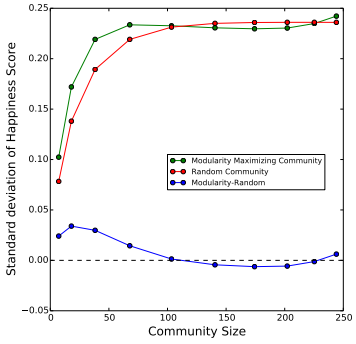
(b)



(c)



(d)



(e)

Figure 11: The standard deviation of happiness of a community as a function of its size. **(a) threshold =15%**, **(b) threshold =20%**, **(c) threshold =25%**, **(d) threshold =40%**, **(e) threshold =50%**. Each point represents a single community. The green points represent communities detected by the modularity maximizing algorithm, the red points represent the random communities, and the blue points represent the difference between the values of respective green and red points.

Set	Description	Cardinality
[A]	raw tweets	478,000
[B]	tweets from [A] mapped inside the grid	404,000
[C]	tweets from [A] in English or Italian	382,000
[D]	tweets from [C] with happiness score $= 0$	55,000
[E]	tweets from [C] with happiness score > 0	327,000
[F]	tweets at the intersection of [B] and [E]	274,000
[G]	cells with happiness score > 0 (using tweets from [F])	5,580
[H]	cells from [G] with ≥ 10 tweets	2,321
[J]	cells from [G] with ≥ 10 unique users	1,213
[K]	cells from [J] within top/bottom 15% of happiness	363
[L]	links with weight ≥ 0.1 including self-edges	2,343,000
[M]	links from [L] with weight ≥ 0.1 without self-edges	2,338,000
[N]	links from [M] with weight ≥ 0.1 (connecting the cells in [K])	52,000

Table 12: Number of tweets, cells and links during each stage of preprocessing.

References

[1] 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.

[2] Jonathan Haynes and Igor Perisic. Mapping search relevance to social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, page 2. ACM, 2009.

[3] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[4] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[5] Mark EJ Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

[6] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[7] Josep M Pujol, Vijay Erramilli, and Pablo Rodriguez. Divide and conquer: Partitioning online social networks. *arXiv preprint arXiv:0905.4918*, 2009.