Under consideration for publication in Network Science

Feature-Based Classification of Networks: Supplementary Materials

Ian Barnett∗
University of Pennsylvania, Department of Biostatistics, Philadelphia, 19104, USA

Nishant Malik∗
Rochester Institute of Technology, Rochester, 14623, USA

Marieke L. Kuijjer
Dana Farber Cancer Instutute, Boston, 02115, USA

Peter J. Mucha
University of North Carolina, Department of Mathematics, Chapel Hill, 27599, USA

Jukka-Pekka Onnela
Harvard University, Department of Biostatistics, Boston, 02115, USA

(*e-mail:* ibarnett@pennmedicine.upenn.edu, nxmsma@rit.edu, mkuijjer@jimmy.harvard.edu, mucha@unc.edu, onnela@hsph.harvard.edu, (617)495-1000)

S1 Materials and methods

S1.1 Sample networks and network feature extraction

In this study, we have used three data sets: call detail records (CDR), tumor gene expressions, and a variety of benchmark social network data. Below we describe the sources of these data sets and the processes used to construct networks from these data sets. Feature-based classification is a two-step procedure, regardless of the application. First, we identify and calculate a set of contextually important network features for each network as described below. Second, we split the data into training and testing sets and then feed the features into the classifier of choice. For a detailed schematic, see Fig. S1.

S1.1.1 Call activity social networks

For the social network setting we use the call detail record (CDR) data from the first, second, and fourth quarter of the year 2014 from a European country's leading telecom operator, which at the time the data were collected had a 57% market share. In these networks, undirected edges are placed between any two individuals who communicated with one another either via phone calls or text messages on the given day. We use different features of network structure and properties of network nodes to classify the networks into

∗ These authors contributed equally.

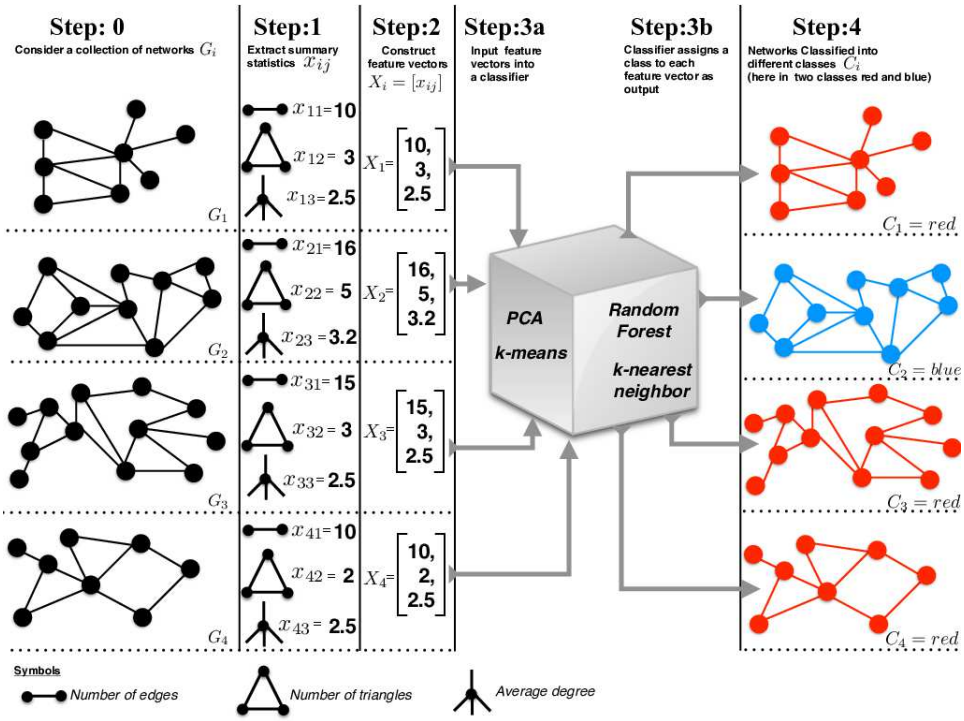


Fig. S1. **A schematic illustrating the steps in our network classification approach**. Here $\{G_i\}$ represents a collection of networks with known class labels, x_{ij} is the jth feature of the ith network, and $X_i = [x_{ij}]^T$ is the (column) feature vector corresponding to the ith network. In principle one can use any of several different classifiers, such as principal component analysis (PCA), k-means clustering, k-nearest neighbor, and random forests.

days of the week and, more generally, into weekday (Monday through Friday) vs. weekend (Saturday and Sunday) networks. Given the natural weekly periodicity in human behavior, we would expect the structure of these networks to reflect changes in the day-by-day social activity and communication patterns. We denote the number of days in the training set with N. Let $X_i = [x_{i1}, \ldots, x_{ip}]^T$ be the p features of the network corresponding to the ith day. National holidays were removed from the analysis because of likely anomalous social behavior on those days.

For each day, the daily call network is constructed by assigning an edge between any two individuals who are in contact by phone on that day. For each day's network, a variety of network features are extracted: the network size (excluding all nodes with degree 0), average clustering coefficient, degree assortativity, fraction of nodes that are female, fraction of edges that are male-female, average age difference over all edge pairs, the fraction of edge pairs from the same zip code, the first four principal components from the degree distribution, and the first four principal components from the clustering coefficient distribution. These features are then used in the selected classifiers to predict whether or not a social / communication network corresponds to a weekend or a weekday, or in the 7-day classifier to a specific day of the week. Table S1 lists all features used in the

weekday/weekend classification. A visualization of the classification results are presented in Fig. 1, with feature importance of the random forest model in Fig. S2.

As indicated above, the features used for classification tasks in the CDR data also includes the four principal components of the degree distribution and of the distribution of the local clustering coefficient. That is, after extracting the daily networks, we compute the distribution of degree $P_i(k_j)$ and local clustering coefficient $P_i(C_j)$ for each network, indexed by i for the day and j for the bin in the respective histogram. For each day, $P_i(k_j)$ and $P_i(C_j)$ distributions includes 80 bins and 20 bins respectively, where $k_j \in [0,80]$ and $C_j \in [0,1]$. There were only few nodes with degree higher than 80, these were removed from the degree distribution to avoid long strings of zeros in the distribution. In place of using the complete set of $P_i(k_j)$ (length 80) and $P_i(C_j)$ (length 20) values as features, we reduce the dimensions (length) using principal component analysis (PCA). Specifically, we treat $P_i(k_j)$ and $P_i(C_j)$ as the input matrices for the PCA and then compute the first four principal components. Table S1 lists the four principal components as DegPC1-4 and ClusPC1-4 while Fig. S3 (a) and (b) plots their values. For comparison, in Fig. S3 (c) we show the first four principal components of the whole feature set used in the classification of CDR data (see Table S1 for the list). We observe in Fig. S3 (a-c) that plots of the first principal components (pc_1) vs. the second principal components (pc_2) distinguishes weekend days and weekdays very well. The first four principal components describe 62% (degree distribution), 90% (local clustering coefficient) and 96% (combined features) of the variances in the respective distributions.

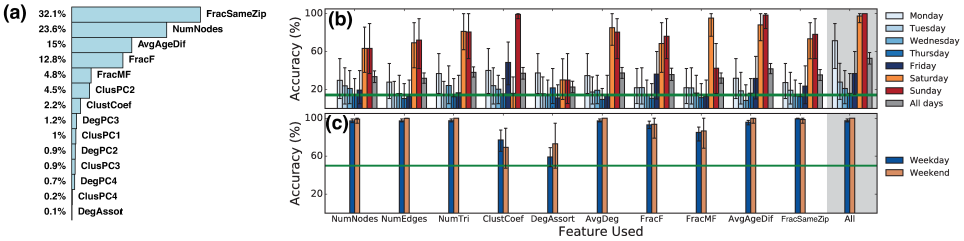


Fig. S2. (a) **Feature importance in the weekend-weekday classification random forest**: Feature importance is calculated from the mean decrease in tree leaf impurity over the full random forest as measured by the Gini index. Percentages are the decrease in impurity for each feature, scaled so they sum to 100%. Three redundant features are not displayed due to their strong correlation with the NumNodes feature. (b) **KNN day-of-week classification**: Each set of bars represents accuracy over multiple realizations of the seven days of the week and the combined accuracy over seven days (grey bars), using a single indicated feature. The last block, highlighted in grey, represents accuracy using all selected features. The green line represents the null rate of classification. (c) **KNN weekend-weekday classification**: Bars represent accuracy of classifying weekdays and weekend days using the indicated features. The green line again represents the null rate.

S1.1.2 Biological networks

Tumor gene expression data was downloaded from The Cancer Genome Atlas for 1217 patients with cancer of the lung (lung adenocarcinoma), brain (glioblastoma multiforme),

or ovary (ovarian serous cystadenocarcinoma). For each sample we reconstructed a bipartite network with edge weights corresponding to the strength of regulation between a transcription factor and a gene, across 10,903 genes and 113 transcription factors (Kuijjer *et al.*, 2015). Given that gene expression levels would be expected to differ by tumor site, we would expect the properties of the bipartite regulatory networks to vary from site to site. In this setting, $N = 547$ is the number of individuals in the training set and $X_i = [x_{i1}, \ldots, x_{ip}]^T$ represent the p features of the ith sample, or individual.

For simplicity, we threshold edge weights in the bipartite network. For each edge, only the top $q\%$ of edge weights across all 1217 networks are declared to be edges, where $q \in [0, 100]$ is the chosen threshold. In other words, for each possible edge, only the networks with edge weights in the top $1217 * (1 - q)$ will be represented as an edge. A large q leads to sparse networks whereas small q leads to dense networks. We use $q = 95$ but to test the sensitivity to q we repeat the analysis using a variety of thresholds.

After the bipartite networks are constructed for each sample, each network is projected onto two unipartite sets, giving gene-gene networks and transcription factor - transcription factor networks. Projection edge weights are defined by the matrix product of the bipartite adjacency matrix with its transpose. Selected features are then extracted from all three network representations. On the bipartite networks, we use average degree, average bipartite clustering coefficient, the mean and variance of node redundancy, and the mean and variance of node closeness centrality. On the unipartite projections, we use the average degree, the number of triangles, average clustering coefficient, and degree assortativity. These features are used to predict what type of cancer tumor the sample was taken from in the remaining 546 samples in the test set. Table S1 lists all features used in the tumor classification.

S1.1.3 Benchmark online and acting social networks

We also compare our approach to six benchmark social network classification tasks previously considered in the literature (Niepert *et al.*, 2016). The online forum Reddit contains many discussion threads about assorted topics. Social networks were constructed for each thread by considering users as nodes and by placing an undirected edge between two users when one had responded to the other in that thread. Some subreddits have more specialized topics. The REDDIT-BINARY data set is used to classify threads as either belonging to a question/answer-based subreddit or a discussion-based subreddit. The REDDIT-MULTI-5K data set contains 5,000 thread networks across five different subreddits and the REDDIT-MULTI-12K data set contains 12,000 thread networks from eleven different subreddits. In both data sets, the aim is to classify a thread into its correct subreddit.

COLLAB is a scientific-collaboration data set, where ego-based networks of researchers from three different fields are constructed with edges to other researchers the ego has collaborated with. The goal is to classify these ego-networks into their correct field.

The IMDB-BINARY data set constructs ego-based networks around every actor where edges are formed between actors that appear in the same movie together. Networks are constructed for two genres, *Action* and *Romance* (ignoring any movie in the union of the

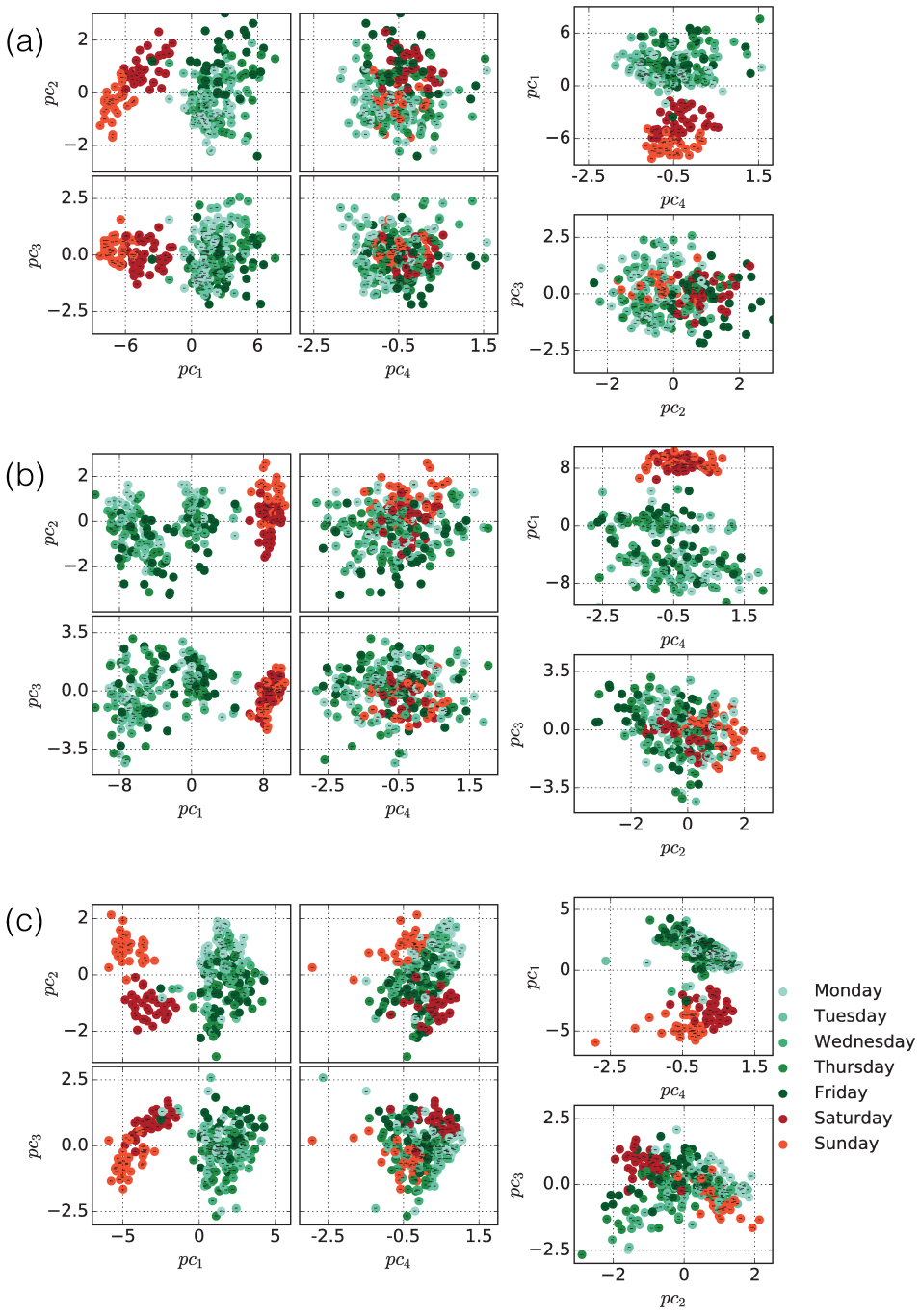


Fig. S3. **Principal Component Analysis (PCA) of the CDR data.** (a) The first four PCA components for local clustering coefficient distributions of daily networks. (b) The first four PCA components for degree distributions of daily networks. (c) The first four PCA components for feature vectors constructed from features listed in Table S1.

Feature label	Feature name
NumNodes	Number of nodes
NumEdges	Number of edges
NumTri	Number of triangles
ClustCoef	Global clustering coefficient
DegAssort	Degree assortativity coefficient(Newman, 2003)
AvgDeg	Average degree
FracF	Fraction of nodes that are female
FracMF	Fraction of edges that are male-female
AvgAgeDif	Average age difference (absolute value) over edges
FracSameZip	Fraction of edges that share the same ZIP code
DegPC1-4	Principal components of degree distribution
ClusPC1-4	Principal components of clustering distribution

Table S1. ***Description of feature labels for weekday/weekend classification.*** *A more detailed description for the feature labels used in Fig. 1, S2, S3 (c) and S4. This includes some redundant features such as NumTri, NumEdges, and DegPC1, all of which are strongly correlated ($\rho > 0.9$) with NumNodes.*

two), with the aim of classifying each ego-network into the correct genre. The IMDB-MULTI data set is similar, but considers three genres, *Comedy*, *Romance*, and *Sci-Fi*.

These six benchmark data sets were previously used to test the classification performance of two graph kernel approaches, Graph Kernels (GK) and Deep Graph Kernels (DGK) (Yanardag & Vishwanathan, 2015), and an approach using convolutional neural networks (PSCN) (Niepert *et al.*, 2016). We compare our approach with the GK, DGK and PSCN results reported in Niepert *et al*. (2016). For each network, we extracted six features to use in our classification: number of nodes, number of edges, average degree, degree assortativity, number of triangles, and the global clustering coefficient. Following the reporting of results in Niepert *et al*. (2016), the accuracy of each of our classifiers was evaluated using 10-fold cross validation.

S1.2 Data driven network classification

S1.2.1 Spatial classifiers: KNN and K-means

For the *KNN* classifier, we start with a training set $\mathcal{T}$ containing the feature vectors $X_i = [x_{i1},\ldots,x_{ip}]^T$, where p is the number of features in the ith sample. Each feature vector X_i in $\mathcal{T}$ is preassigned a known class $Y_i \in \{1,\ldots,c\}$. These classes could be days of the week as in CDR data set or disease sites as in the cancer data set. We find the k-nearest neighbors of a new feature vector X_j in the prediction set $\mathcal{P}$ using Euclidean distance $d(X_i,X_j) = \sqrt{\sum_{l=1}^{l=p}(x_{il}-x_{jl})^2}$ and classify it into the Y_j class, $Y_j \in \{1,\ldots,c\}$ by the majority vote among the k-nearest neighbors.

In contrast, k-means clustering provides an unsupervised classification system, wherein one partitions the complete set of feature vectors $\{X_i\}_{i=1}^N$ into a set $C = \{C_1,C_2,\ldots,C_k\}$ of $k \leq N$ clusters. These clusters are found by minimizing the square of the distance from the data points X_j to the center of a cluster i.e., solving

$$\arg\min_{C_i} \sum_{i=1}^{k} \sum_{X_j \in C_i} (||X_j - \mu_i||)^2,$$

where $\mu_i,\ i = 1,\ldots,k$ is the position of cluster C_i. After this clustering, available known classification properties within each cluster can be assessed by various measures. Fig. S4 shows an output of the k-means clustering for the CDR data, classifying Saturdays, Sundays and the weekdays.

We refer the reader to Friedman *et al.* (2001) for a far more complete description of different methods for classifying and assessing classifications.

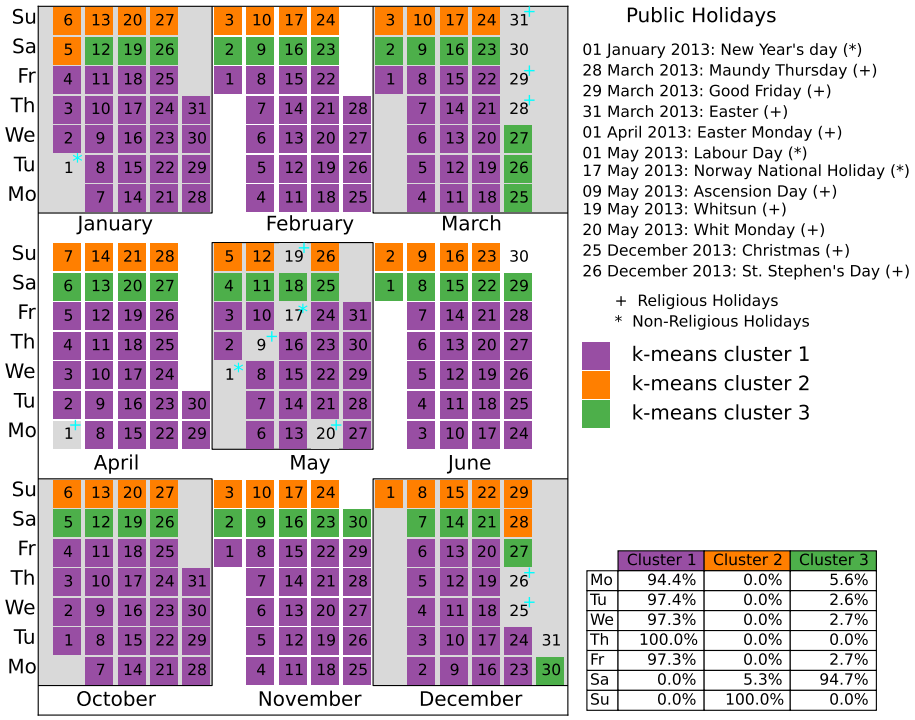
S1.2.2 Random forest classifier

With p features extracted, classification trees (Friedman *et al.*, 2001) can be used to identify the subset of features that are important in distinguishing classes of networks (i.e. weekend days from weekdays). To begin construction of a tree, the data is split into the two groups that best separate the classes. Specifically, let

$$R_1(j,s) = \{X_i : x_{ij} < s\} \text{ and } R_2(j,s) = \{X_i : x_{ij} \geq s\} \tag{1}$$

be the regions that separate the data into two groups. Consider, for example, the classification of CDR social networks into weekend days and weekdays. Letting $\hat{p}_k$ be the fraction of data points in region R_k that are weekdays, the regions that best separate the weekend days from the weekdays are determined by minimizing $\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)$ with respect to j and s. The $\hat{p}_k(1-\hat{p}_k)$ function here is known as the *Gini index*, penalizing the kth region if $\hat{p}_k$ is far from 0 or 1, as this indicates that the region does not separate the weekdays from weekend days very well. The minimization of the *Gini index* only considers each individual branch of the classification tree at a time.

This process is repeated on the two resulting branches. This is repeated further until the data has been split too many times and there is only one data point in one of the branches, at which point the splitting on that branch terminates. For the minimization occuring at each branching, the random forest approach is as described above except we only consider



	Cluster 1	Cluster 2	Cluster 3
Mo	94.4%	0.0%	5.6%
Tu	97.4%	0.0%	2.6%
We	97.3%	0.0%	2.7%
Th	100.0%	0.0%	0.0%
Fr	97.3%	0.0%	2.7%
Sa	0.0%	5.3%	94.7%
Su	0.0%	100.0%	0.0%

Fig. S4. **Classification of days from daily call record data using k-means clustering.** The feature vector in this case was composed of features listed in Table S1. Holidays were removed from the data prior to running the classification routine.

a random subset $m \leq p$ of the features (we use $m = 4 \approx \sqrt{p}$) while also bootstrapping the data at each branching. This introduces randomness into the tree building process, and $B = 10,000$ such random classification trees are built. In order to classify a new data point, x, let $\hat{C}_b(x)$ be the class prediction of the bth random forest tree. The classification of x is determined to be the majority vote over all $\{\hat{C}_b(x)\}_1^B$.

To apply this procedure to more than two outcomes (seven day classification as opposed to weekend versus weekday), the procedure is similar except the Gini index becomes $\sum_{k=1}^{7} \hat{p}_{rk}(1-\hat{p}_{rk})$ where $\hat{p}_{rk}$ represents the proportion of data points in region r that represent day of the week k. The classification tree is built using odd days over all three available quarters, and the model is tested on the even days.

This same approach can applied to the biological network context to predict tumor type. For tumor type we consider three possible outcomes: brain, lung, and ovary. In our implementation, p is the same so again we use $m = 4 \approx \sqrt{p}$.

S2 Further exploration of CDR data set

In this section, we discuss further details of the CDR data. We analyze some of the structural properties of the extracted networks and discuss network sampling.

S2.1 Degree Distribution

To further identify differences between the days of the week, we have estimated the degree distributions $p(k)$ for individual days, fitting them to lognormal distributions:

$$p(k) = \frac{1}{k\sigma\sqrt{2\pi}} e^{\frac{-(\ln k-\mu)^2}{2\sigma^2}}. \tag{2}$$

Table S2 provides the values of the fitted parameters and their estimated confidence intervals, whereas in Fig. S5 we have plotted the empirical and fitted degree distributions. In particular, we observe in Fig. S5 that weekends appear to have distinct distributions from the weekdays. The parameters of the fitted distributions are similarly distinct for weekdays compared to weekends (see Table S2).

S2.2 Distribution of Age and Clustering

Age plays a significant role in the way people use mobile phones, hence we might expect that communication patterns are influenced by user age. Average age difference across communication ties emerged as an important feature for the CDR data set (see Fig. S2). In Fig. S6, we plot the distribution of age in the data and the corresponding average clustering in the network corresponding to a particular age group. One of the striking features here is that weekend and weekday networks show distinct patterns for average clustering versus age. We also observe rather higher clustering for ages below 20, probably implying that these users interact within small tightly knit local network neighborhoods. In addition, most communication occurs between individuals of similar age, while there also appears to be a generational gap in high frequency communication between people approximately 25 years apart (see Fig. S7), which likely reflects parent-child communication.

S2.3 Network sampling

The accuracies obtained above and in the main text are due, in no small part, to the large quantity of available data. However, in many scenarios the study design does not allow the luxury of the full network from such a massive sample size. In this case, one must use a subsample of the network. To investigate the effect of network subsampling on predictive power, we compare the performance of two subsampling procedures, sampling on geography and snowball sampling. Sampling on geography (via ZIP codes) selects a subset of individuals who live in close proximity to one another, whereas snowball sampling starts with a seed node in the network and branches out from that node following its edges, going several edges away from the seed node, recruiting the nodes along the path to the sample.

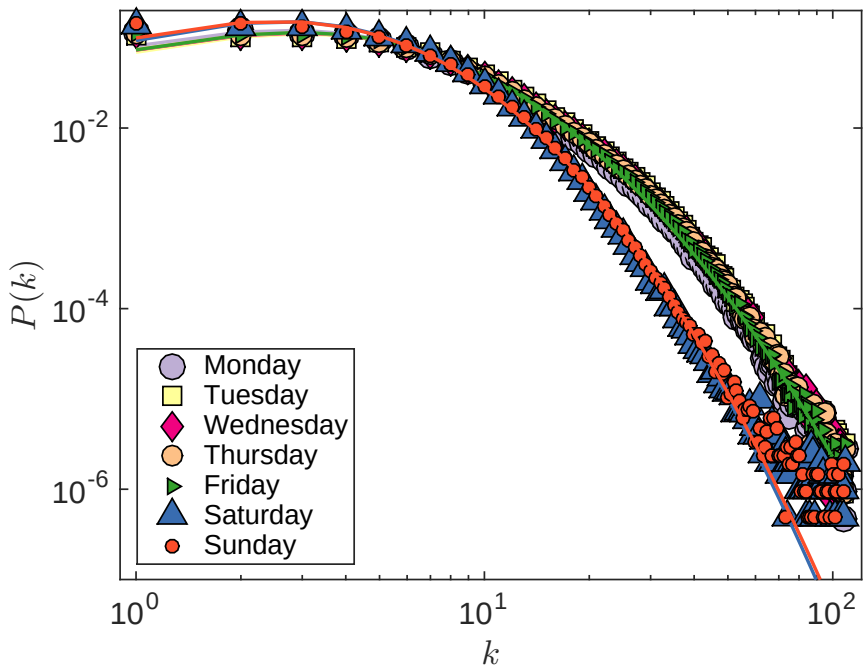


Fig. S5. **Degree distribution for each day of the week.** Thick lines are fitted distributions given by Eq. 2, with values of the fitted parameters given in Table S2.

Day	μ	σ	p-value
Monday	1.794 ± 0.004	0.693 ± 0.003	0.656
Tuesday	1.885 ± 0.004	0.710 ± 0.003	0.449
Wednesday	1.860 ± 0.004	0.710 ± 0.003	0.989
Thursday	1.853 ± 0.004	0.708 ± 0.003	0.811
Friday	1.851 ± 0.004	0.702 ± 0.003	0.709
Saturday	1.650 ± 0.004	0.603 ± 0.003	0.360
Sunday	1.636 ± 0.004	0.613 ± 0.003	0.498

Table S2. ***Values of fitted log-normal parameters and*** 95% ***confidence intervals.*** *The confidence intervals were constructed assuming the asymptotic normality of the maximum likelihood estimate. The p-values are obtained employing two-sample Kolmogorov-Smirnov tests under the null hypothesis that the fitted distribution and the sample distribution are the same continuous distribution. The test indicates that the fitted and sample distributions can not be statistically distinguished from one another.*

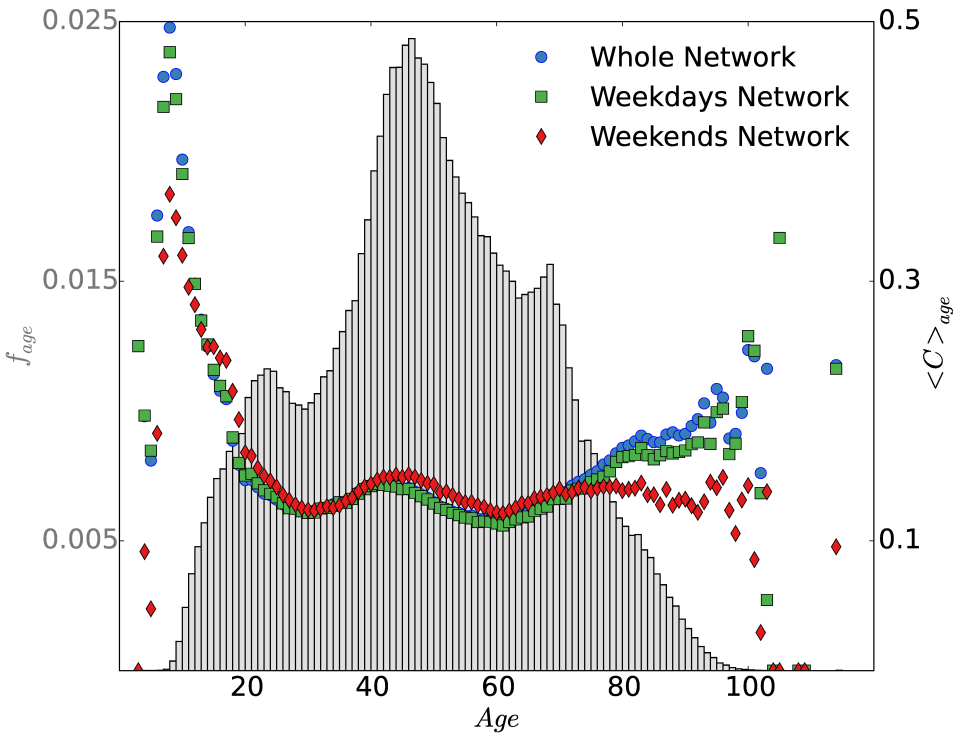


Fig. S6. **Distribution of age and average clustering.** Here we consider three networks, one built from the whole data set across the three quarters of the year (whole network), one built across the data set but limited to week days (weekdays network), and finally one built across the data set but limited to weekends (weekends network). The grey histogram gives the distribution of age in the CDR data set. Blue dots are the average local clustering coefficients for nodes with the given age for the whole network data, whereas green squares and red diamonds represent the average clustering coefficients for the weekday and weekend networks, respectively.

We fit the following model:

$$\frac{1}{MR_i + \delta} = \beta_0 + \beta_1 X_i + \beta_2 Z_i X_i + \varepsilon_i \tag{3}$$

where MR_i is the misclassification rate based the ith subsample, X_i is the average daily network size (number of nodes) based on the ith subsample, Z_i is an indicator for if the ith observation was based on a ZIP code subsample, $\delta = 0.01$ is a shift used to avoid division by 0, and $E[\varepsilon_i] = 0$ with $\text{Var}(\varepsilon_i) = \sigma_i^2$. In addition, we force $\beta_0 = (5/7 + \delta)^{-1}$ to reflect the fact that when sample size is 0 the misclassification rate is $5/7$ (corresponding to the classifier that predicts every day to be a weekday). Due to strong heteroskedastic errors in this model and a strong presence of outliers, model (3) is fit using least-absolute-deviations regression (Barrodale & Roberts, 1973). We test for a difference in the misclassification rates when comparing snowball to ZIP code sampling. This test corresponds to the hypotheses $H_0 : \beta_2 = 0$ and $H_A : \beta_2 \neq 0$. To perform inference on $\hat{\beta}_2$ we simulate the null distribution of $\hat{\beta}_2$ by permuting sampling-type labels (Z_i). Due to ZIP code and snowball samples having different distributions in average network size (ZIP code samples tend to be

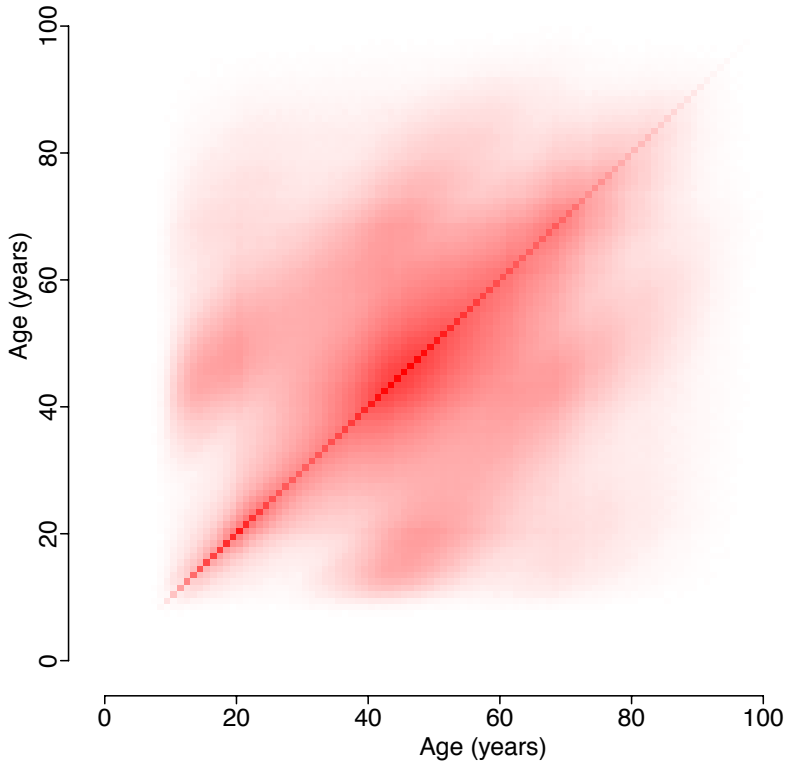


Fig. S7. **Frequency of communication between age groups.** The full social network from the combined Q1, Q2, and Q4 of 2013 is used to count the number of age-age network edges. High frequency age-age connections are dark red, with less common age-age pairings in light red. The dark diagonal corresponds to communications within the same age group. The red intensity of the (i, j) cell is $(x_{ij}/\max_{ij} x_{ij})^4$ where x_{ij} is the total number of edges people of age i share with people of age j.

larger), we put observations in bins of width 20 and permute Z_i labels only on observations within each bin. This preserves the distribution of average network size amongst both types of sampling procedures.

The misclassification rate of the weekend/weekday random forest classifier for all considered snowball and ZIP code subsamples are displayed in Fig. S8. After fitting model (3), which relates network size and subsampling procedure to misclassification rate, we found that $\hat{\beta}_2 = -0.03$. This implies that when holding the network size fixed, the slope of the expected misclassification rate on the inverse scale is -0.03 lower for ZIP code sampling than it is for snowball sampling. This change in slope is significant (p-value $= 6.5 \cdot 10^{-5}$), implying that, on average, snowball sampling yields networks that have features that inform classification of weekends and weekdays better than those from ZIP code samples of equivalent sample size.

There is a clear tendency for the ZIP code subsamples to have higher misclassification rates relative to the size of the subsample than is the case for snowball subsamples. A large part of the reason for this greater misclassification in ZIP code subsamples could be because the feature measuring the fraction of ties that are within the same ZIP code holds no meaning for ZIP code subsamples (trivially the fraction is always one). As seen in Fig. S2, this feature is vitally important to classification of weekends from weekdays, so ZIP code subsamples suffer without it.

To perform a snowball sample based on a given day's network, a random seed node is selected and all nodes within a distance of 4 are included in the subsample. This is repeated for each day in the data set as well as for radii of 5 and 6. In each case, if the resulting network has fewer than 50 nodes, a new random seed node is selected until the subsample of sufficient size is acquired. ZIP code subsamples include all individuals from the same ZIP code regardless of their social connections. Each ZIP code in the country matching with at least 50 active customers in the data set was used as a separate subsample. Altogether there were 247 snowball subsamples and 293 ZIP code subsamples included in the analysis. The same binary random forest classification procedure was replicated for each subsample as was performed on the full data set, and the resulting misclassification rates for each subsample were recorded.

References

Barrodale, Ian, & Roberts, Frank DK. (1973). An improved algorithm for discrete l_1 linear approximation. *Siam journal on numerical analysis*, **10**(5), 839–848.

Friedman, Jerome, Hastie, Trevor, & Tibshirani, Robert. (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.

Kuijjer, Marieke Lydia, Tung, Matthew, Yuan, GuoCheng, Quackenbush, John, & Glass, Kimberly. (2015). Estimating sample-specific regulatory networks. *arxiv preprint arxiv:1505.06440*.

Newman, Mark EJ. (2003). Mixing patterns in networks. *Physical review e*, **67**(2), 026126.

Niepert, Mathias, Ahmed, Mohamed, & Kutzkov, Konstantin. (2016). Learning convolutional neural networks for graphs. *arxiv preprint arxiv:1605.05273*.

Yanardag, Pinar, & Vishwanathan, SVN. (2015). Deep graph kernels. *Pages 1365–1374 of: Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*. ACM.

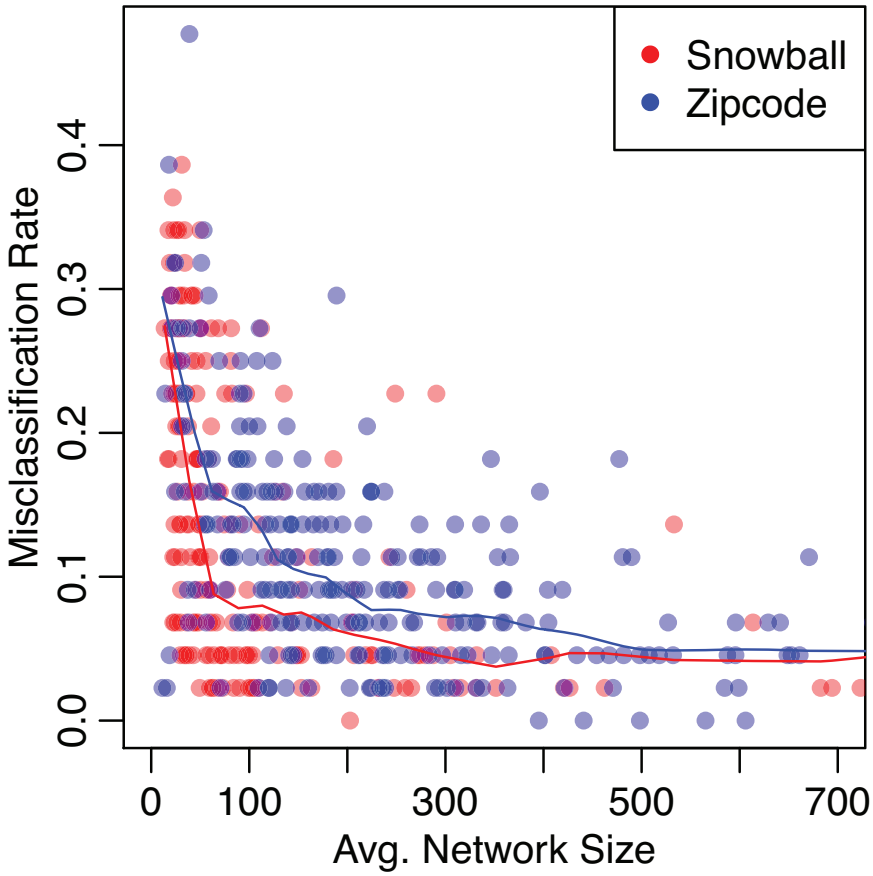


Fig. S8. **Misclassification rates of classifiers built on network subsamples.** Each point represents a random forest classifier built from a subsample of the full network. Networks are constructed on a daily basis. The inverse-transformed misclassification rate is used to fit the regression model. Snowball samples with radii 4, 5, and 6 are included. Each subsampled network has varying network size on any given day because the networks are constructed using edge lists, so if a node in the original subsampled network has degree 0 on a particular day, then that node will not appear in that day's network. A smooth is fit to each point cloud, one for snowball subsamples and one for ZIP code subsamples.