SCIENTIFIC REPORTS

OPEN

Understanding tie strength in social networks using a local "bow tie" framework

Received: 19 December 2017
Accepted: 3 May 2018
Published online: 19 June 2018

Heather Mattie[1], Kenth Engø-Monsen[2], Rich Ling[3] & Jukka-Pekka Onnela[1]

Understanding factors associated with tie strength in social networks is essential in a wide variety of settings. With the internet and cellular phones providing additional avenues of communication, measuring and inferring tie strength has become much more complex. We introduce the social bow tie framework, which consists of a focal tie and all actors connected to either or both of the two focal nodes on either side of the focal tie. We also define several intuitive and interpretable metrics that quantify properties of the bow tie which enable us to investigate associations between the strength of the "central" tie and properties of the bow tie. We combine the bow tie framework with machine learning to investigate what aspects of the bow tie are most predictive of tie strength in two very different types of social networks, a collection of medium-sized social networks from 75 rural villages in India and a nationwide call network of European mobile phone users. Our results show that tie strength depends not only on the properties of shared friends, but also on non-shared friends, those observable to only one person in the tie, hence introducing a fundamental asymmetry to social interaction.

The strength of any kind of relationship between two individuals lies on a spectrum. People in general have a close relationship with only a few friends or family members, a somewhat weaker tie with a larger group of individuals with whom they interact less frequently, and an even weaker connection with a large number of casual acquaintances. This tradeoff between tie strength and the number of people a person is connected to through his or her ties was elegantly captured by Dunbar[1]. Measuring and predicting tie strength, and moreover, understanding the factors that drive tie strength, has been an expanding area of interest, with increasing utility and complexity in the digital age, i.e., the ever-increasing forms of communication via mobile phones and social media. Knowledge of the strength of a tie, as well as the social dynamics contributing to tie strength, has been shown to increase the accuracy of link prediction, enhance the modeling of the spread of disease and information, and lead to more targeted marketing[2–4].

Several indicators of tie strength have been proposed, perhaps most notably by Mark Granovetter in his seminal work The Strength of Weak Ties[5]. Granovetter differentiated between strong and weak ties and proposed the weak ties hypothesis: the stronger the tie between any two people, the higher the fraction of friends they have in common[5]. Much of the current methodology centered on tie strength has stemmed from Granovetter's weak ties hypothesis and his proposed four dimensions of tie strength: the amount of time spent interacting with someone, the level of intimacy, the level of emotional intensity, and the level of reciprocity. More recently, three additional dimensions of tie strength have been proposed: (1) emotional support[6,7], (2) structural variables, i.e. network topology[8–10], and (3) social distance, i.e. the difference in socioeconomic status, education level, political affiliation, race, and gender[9,11]. These categories have facilitated the definition and quantification of numerous possible predictors of tie strength; some generalizable to any network, and some specific to a limited number of social networks.

Another hypothesis of importance to this analysis is a corresponding perspective outlined by Elizabeth Bott[12] that suggests that the tie strength between husband and wife varies *inversely* with the number of non-overlapping ties. That is, overlapping (common) friends support the tie strength between husband and wife, and non-overlapping friends, i.e. friends in each spouse's separate social circle, detract from it. Several studies have tested Bott's hypothesis with mixed findings. The studies that did not find evidence to support the hypothesis suffer from non-representative samples, a lack of statistical analysis, and confounding from age, social class and gender[7,13–15].

[1]Harvard T.H. Chan School of Public Health, Biostatistics, Boston, 02115, USA. [2]Telenor Research, D4d, Snarøyveien 30, Fornebu, N-1360, Norway. [3]Wee Kim Wee School of Communication and Information, Nanyang Technological University, 31 Nanyang Link, Singapore, 637718, Singapore. Correspondence and requests for materials should be addressed to H.M. (email: hemattie@hsph.harvard.edu)

Initially, highly generalizable similarity indices such as the number of common neighbors two nodes share, preferential attachment, and path distance were used to infer tie strength. These metrics were most commonly used for link prediction and were shown to provide some information regarding tie strength[3,16]. However, it was quickly discovered that the addition of nodal attributes and other metrics not solely based on network topology greatly enhanced the measurement and prediction of tie strength[17,18]. Gilbert and Karahalios defined indicators of tie strength specific to a network of Facebook users and built a predictive model that achieved 85% accuracy for binary tie strength (weak vs. strong) classification[19]. They found that the act of communicating once leads to a significant increase in tie strength, and that educational difference plays a role in determining tie strength. Pappalardo *et al*. introduced a measure of tie strength using multiple online social networks and found that the strength of a tie is related to the number of interactions between the two individuals[16]. In addition, several studies have shown that frequent communication, both online and offline, is positively related to tie strength[6,20].

While previous studies have provided advances and valuable insights, they suffer from a binary definition of tie strength (weak vs strong), low diversity in the types of social networks studied (the vast majority being social media sites), and non-representative samples. In this work, we propose a decomposition of a social network into an ensemble of interconnected "social bow ties," constellations consisting of nodes and ties that surround each network tie. We call any such subgraph a "social bow tie" because the topological structure that surrounds each tie resembles a bow tie. We also introduce several simple metrics that quantify properties of the bow tie. Further, we use random forests and linear regression to build models that predict categorical and continuous measures of tie strength from different properties of the bow tie, including nodal attributes (covariates) of the nodes included in the bow tie. We apply our framework to two social networks, a collection of 75 social networks from the villages of Karnataka, India, and a call network of European mobile phone subscribers. We find that the bow tie framework contributes to more accurate predictions of tie strength and provides insights on which metrics are the most informative of tie strength. Specifically, we find that the larger the proportion of shared friends, the stronger the tie, and the more clustered the individual friendship circles (consisting of non-overlapping friends), the weaker the tie. Consequently, these findings provide evidence to support both the weak ties hypothesis and a generalized version of the Bott hypothesis[12].

Methods

Data Description. We analyzed two social network data sets. The first data set is social network data collected in 2006 from 75 villages located in 5 districts in rural southern Karnataka, India. The data were collected through household and individual surveys as part of a study by Banerjee *et al.*[21]. Of relevance for this study, the survey included social network data along 12 dimensions: friends or relatives who visit the respondent's home, friends or relatives the respondent visits, any kin in the village, non-relatives with whom the respondent socializes, those from whom who the respondent receives medical advice, with whom who the respondent goes to temple to pray, from whom the respondent would borrow money, to whom the respondent would lend money, from whom the respondent would borrow material goods from, to whom the respondent would lend material goods, from whom the respondent gets advice, and to whom the respondent gives advice. It is worth noting that these forms of interaction are largely face-to-face, unlike the mediated material from the call detail records (CDRs) described below. Additionally, a proportion of villagers were given individual surveys that recorded age and sex, among other attributes.

For this data set, we define the strength of a tie as the number of distinct types of social relationships reported to exist between the two individuals. For example, if individual i borrows money from individual j and in addition gives advice to individual j, the weight of the (undirected) tie between i and j would be equal to 2. If i and j also attend temple together, their tie strength would be 3 and so on, with a minimum strength of 1 and a maximum strength of 12 for any tie. Note that a tie strength of 0 implies that the two individuals are not connected by any kind of social tie. We denote the strength of a tie between individuals i and j as w_{ij}. Because we ignore the directionality of ties, our definition of tie strength is symmetric.

The second data set consists of call detail records (CDRs) from a mobile phone provider in an undisclosed European country where 68% of citizens own a smartphone and 85% own a cellular phone. The data examined here span a period of three months in 2013, and each record consists of the following daily aggregate communication summaries for pairs of individuals: the date, anonymized caller ID, anonymized callee ID, daily call duration (in minutes), daily number of calls, daily number of text messages (SMS), and daily number of multimedia messages (MMS). Age, sex, and billing zip codes were available for a large majority of individuals.

An undirected, weighted call network was created from the records by first summing the call durations between any two individuals over the three-month period. If two individuals spoke on the phone at least once during the period, we connected them with an edge of strength w_{ij}, where the value of edge strength was set to the total amount of time spent on the phone with one another. Since tie strength is defined in terms of absolute time, it does not take into account the total amount of time each individual spends on the phone, which makes it somewhat difficult to quantify the relative strength of ties since the strength of a tie is not measured on the same scale either for individuals or pairs of individuals. We therefore normalized tie strength and represent it with two measurements: one that represents tie strength from the perspective of individual i, and one that represents tie strength from the perspective of individual j. Specifically, for each tie, the first measurement of tie strength is the total call duration (w_{ij}) divided by the total time individual i spends on the phone s_i, the strength of node i. similarly, the second measurement of tie strength is the total call duration divided by the total time individual j spends on the phone s_j, the strength of node j. Dividing total call duration by the strength of each focal node results in a consistent definition of tie strength. We denote these new tie strength measurements as y_{ij} and y_{ji}. We created another summary measure of tie strength by taking the average of y_{ij} and y_{ji}, and we denote this $z_{ij} = (y_{ij} + y_{ji})/2$.

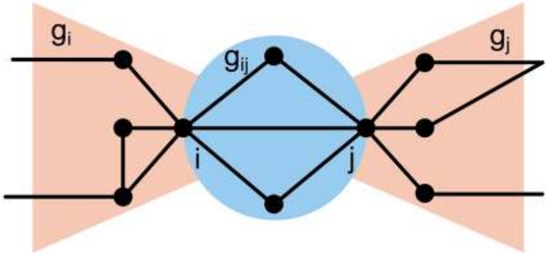

Figure 1. A simple example of the social bow tie G_{ij}. The blue circle contains the nodes and edges that comprise the overlapping friendship circle of the focal nodes *i* and *j*, denoted g_{ij}. The parts of the bow tie shaded in orange contain the individual (non-overlapping) social circles of the focal nodes, denoted g_i for node *i* and g_j for node *j*.

Bow Tie Framework. To introduce the "bow tie" structure, consider a weighted social network *G*, which may be directed or undirected, and consider a tie with weight w_{ij} that connects two individuals *i* and *j*. We call these two individuals the *focal nodes* of the bow tie. We use the term *focal tie* to refer to the tie that links them. We start by partitioning *i*'s friends and *j*'s friends into three disjoint sets. Group *i*, denoted g_i, contains the nodes that are connected to only *i*; group *j*, denoted g_j, contains nodes that are connected to only *j*; and group *ij*, denoted g_{ij}, contains nodes that are connected to both *i* and *j*. These three groups jointly make up the shared and non-shared friends of *i* and *j*. We call this structure the *ij bow tie*. Formally, the groups g_i, g_j and g_{ij} are induced subgraphs, where the node sets that induce them are the neighbors of *i*, the neighbors of *j*, and the common neighbors of *i* and *j*, respectively. The bow tie *ij*, denoted by G_{ij}, is the subgraph that is induced by the union of all neighbors of *i* and *j*. Note that G_{ij} is more than the sum of g_i, g_j and g_{ij}: in addition to containing the same set of nodes and ties as those subgraphs do, it also contains the inter-group ties among this set of nodes, i.e., the ties linking nodes across g_i, g_j and g_{ij}. Important to our analysis below is the hierarchical structure of the bow tie: at the upper level of hierarchy we have the bow tie G_{ij}; at the intermediate level, we have the three groups, g_i, g_j and g_{ij}; and at the lowest level we have the nodes and ties from which each group is composed. A simple example of the bow tie structure surrounding nodes *i* and *j* is shown in Fig. 1. While we were inspired by the well-known WWW topology bow tie structure presented by Broder *et al.*[22], the framework introduced here is quite different. Broder *et al.* view the internet at a global, macroscopic level, while the social bow tie is a local, microscopic structure.

The localized nature of the bow tie framework gives rise to several topological metrics that can be used to predict tie strength and find evidence for or against both the weak ties hypothesis and the Bott hypothesis. We include unweighted[23] and weighted[24] edge overlap, which we denote o_{ij} and $\tilde{o}_{ij}$, respectively. Unweighted overlap is defined as in (1), and weighted overlap as in (2).

$$o_{ij} = \frac{n_{ij}}{k_i + k_j - 2 - n_{ij}} \tag{1}$$

$$\tilde{o}_{ij} = \frac{\sum_{k=1}^{n_{ij}}(w_{ik} + w_{jk})}{s_i + s_j - 2w_{ij}} \tag{2}$$

Here, n_{ij} is the number of common (shared) friends of nodes *i* and *j*, k_i (k_j) denotes the degree, or number of connections, node *i* (*j*) has, w_{ij} denotes the weight associated with the tie between nodes *i* and *j*, and s_i (s_j) denotes the strength of node *i* (*j*). In accordance with the weak ties hypothesis, we expect both o_{ij} and $\tilde{o}_{ij}$ to be positively associated with tie strength, i.e., that tie strength w_{ij}, increases as the number of shared friends increases. Metrics based on customized versions of the clustering coefficients of *i* and *j* are used, where the calculation of a clustering coefficient is limited to the non-shared friends of each node, i.e., for node *i*, the nodes and edges in g_i are used to calculate the clustering coefficient of *i*, and similarly, g_j is used for node *j*. We denote the sum and absolute difference of these quantities as cc_{ij}^S and cc_{ij}^D for the unweighted clustering coefficients, and $\tilde{c}c_{ij}^S$ and $\tilde{c}c_{ij}^D$ for the weighted clustering coefficients. Here, we use the definition of weighted clustering coefficient provided by Saramäki *et al.*[25]. Specifically, the weights of ties are considered and the metric reflects how large triangle weights are compared to a network maximum. Other predictors include the sum and absolute difference in the degrees of *i* and *j* (k_{ij}^S and k_{ij}^D), the sum and absolute difference in the strengths of *i* and *j* (s_{ij}^S and s_{ij}^D), the number of nodes and edges in g_{ij} (n_{ij} and e_{ij}), and the sum and absolute difference in the number of nodes and the number of edges in g_i and g_j (n_{ij}^S, n_{ij}^D, e_{ij}^S and e_{ij}^D). With these definitions, we can represent a generalized version, i.e. one that applies to all ties in the network, of Bott's hypothesis in two different ways; using s_{ij}^S and cc_{ij}^S. Bott suggests that the more close-knit the non-overlapping social circles of two connected individuals, the weaker the tie between them. Translating this to our setting, we expect tie strength to be negatively associated with s_{ij}^S and cc_{ij}^S. Specifically, as the clustering and strength of ties among individuals in g_i and g_j increases, tie strength (w_{ij}) decreases. Finally, predictors created from the attributes of *i* and *j* include the sum and absolute difference in the ages of *i* and *j* (a_{ij}^S and a_{ij}^D), the paired sex category (male-male, female-female, female-male) denoted I_{MM}, I_{FF} and I_{FM} respectively, and an indicator if *i* and *j* have the same billing zip code, denoted Z_{ij}. See Table 1 for a detailed description of each variable.

To predict tie strength and study how it is associated with different metrics, we used regression as well as Random Forest (RF) regression and classification[26]. For the India social network, tie strength is discrete with

Predictor	Description
k_{ij}^{S}	Sum of the degrees of i and j ($k_i + k_j$)
k_{ij}^{D}	Absolute difference in the degrees of i and j ($\lvert k_i - k_j \rvert$)
s_{ij}^{S}	Sum of the strengths of i and j ($s_i + s_j$)
s_{ij}^{D}	Absolute difference in the strengths of i and j ($\lvert s_i - s_j \rvert$)
cc_{ij}^{S}	Sum of the clustering coefficients of i and j
cc_{ij}^{D}	Absolute difference in the clustering coefficients of i and j
$\tilde{c}c_{ij}^{S}$	Sum of the weighted clustering coefficients of i and j
$\tilde{c}c_{ij}^{D}$	Absolute difference in the weighted clustering coefficients of i and j
a_{ij}^{S}	Sum of the ages of i and j
a_{ij}^{D}	Absolute difference in the ages of i and j
Sex_{ij}	Categorical variable indicating a male-male, female-female, or female-male tie
I_{MM}	Indicator variable of a male-male tie
I_{FF}	Indicator variable of a female-female tie
I_{FM}	Indicator variable of a female-male tie
Z_{ij}	Indicator if i and j have the same billing zip code
o_{ij}	Unweighted overlap of edge between i and j
$\tilde{o}_{ij}$	Weighted overlap of edge between i and j
n_{ij}	Number of common friends of i and j
e_{ij}	Number of edges among the common friends of i and j
n_{ij}^{S}	Sum of the number of nodes in g_i and g_j
n_{ij}^{D}	Absolute difference in the number of nodes in g_i and g_j
e_{ij}^{S}	Sum of the number of edges in g_i and g_j
e_{ij}^{D}	Absolute difference in the number of edges in g_i and g_j

Table 1. Descriptions of tie strength predictors.

$w_{ij} \in \{1, \ldots, 12\}$. Thus, the weight of a tie can be viewed as a categorical outcome, allowing RF classification and Poisson regression to be used to predict tie strength, or as continuous with RF regression used for prediction. For the CDR call network, tie strength is most naturally treated as a continuous variable, and we used RF regression and linear regression to predict both measures of tie strength.

In addition to ordinary least squares (OLS) regression, least absolute shrinkage and selection operator (LASSO) and ridge regression were used to fit more parsimonious and interpretable models as well as increase prediction accuracy. Before using LASSO and ridge regression, all data was centered around the mean and 10-fold cross validation was performed to select the best tuning parameters; denoted λ^{L} for LASSO and λ^{R} for ridge regression. For RF classification, the number of trees used was 200, and the maximum number of features (covariates) considered when splitting a node was $\sqrt{n}$ where n is the total number of features. For RF regression, 200 trees were used and the maximum number of features considered when splitting a node was n.

Data availability. The India social network data analyzed during the current study are available in the Harvard Dataverse repository, https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21538. CDR data that support the findings of this study are available from Telenor, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Results

India Social Network.

The India network contained 69,444 nodes, of which 16,984 (24.5%) had full attribute information available, and 294,778 edges after the removal of isolated ties. Of these, 37,714 (12.8%) edges were between two individuals with complete attribute information available. The amount of nodal attribute missingness in the India network was high, 75.5%, and we therefore determined that imputation might significantly impact the results, and decided not to impute nodal attributes for this data set. This is because imputation methods for network (correlated) data are not yet fully developed. Consequently, we only included node pairs that had no missing attributes as focal ties. However, all nodes and ties contained in the bow tie structure surrounding each focal tie were used in the calculations and analysis. This was possible since attribute information is only needed for the focal nodes, and not the nodes in the surrounding bow tie structure. Thus, the network topology was not disturbed. We discovered tie strength had a bimodal distribution with ≈46% of ties having a maximum strength of 12. This was due to the fact that the majority (96%) of ties between individuals living in the same household had a weight of 12. We decided to exclude ties between individuals from the same household and only included cross-household ties as focal ties. This resulted in a Poisson distribution of tie strength and a total of 21,945 ties. Similar to the reasoning above, including only cross-household focal ties does not disrupt the topology of the network, but rather the generalizability of the results. Excluding within-household ties as focal ties implies our results cannot be applied to within-household ties. However, in this data set, 96% of within-household ties have a tie strength of 12 and were therefore deterministic. Additionally, according to Banerjee *et al.*[21], nodal attributes were collected from all individuals in a household from a random sample of households in each village, and are assumed to be representative of the population.

RF regression and classification were used to fit three models both before and after nodal attribute imputation, where ties with complete attribute information available were included in the analysis before imputation and all ties were included after imputation. Model 1 is the full model and includes all covariates described in Table 1 with the exception of Z_{ij} since it is specific to the CDR data set; Model 2 includes all covariates except weighted overlap; and Model 3 includes all covariates except unweighted overlap. It has been shown that categorical predictors do not need to be split into multiple dichotomous covariates (referred to as dummy variables) when implementing RF if there are a small number of them and their cardinality is low[26,27]. Therefore, the variable *Sex* was not split into two separate dummy variables due to its low cardinality and it being the single categorical predictor. Accuracy was measured as the residual, the absolute difference between empirical tie strength (w_{ij}) and predicted tie strength ($\tilde{o}_{ij}$). Figure 2 shows the accuracy of RF regression and classification for all models. Note that only two lines are visible, one for RF regression and one for RF classification since the accuracy of all models is indistinguishable. Within one unit of tie strength, an accuracy of 36.4% and 55.3% was achieved by RF regression and classification, respectively.

Feature importance for each of the three models for both RF regression and classification is shown in Fig. 2. The horizontal bars represent how informative the predictor is with a longer bar meaning more informative. The black vertical line represents the value of an equilibrium or null importance if every predictor were equally informative. For both classification and regression, weighted overlap ($\tilde{o}_{ij}$) is the most informative variable in models 1 and 3, and the sum of the clustering coefficients (cc_{ij}^{S}) is the most informative in model 2, followed by the sum of the number of friends in the non-overlapping social circles (n_{ij}^{S}). These results provide evidence that the proposed indicators of tie strength in the Weak Ties and Bott hypotheses (the overlap of friendship circles and the amount of clustering in the non-overlapping friendship circles) are predictive of tie strength.

Poisson regression was used to model the associations between tie strength and each of the predictors, and the coefficients of significant predictors with magnitude greater than 0.2 are reported in (3). The predictors with the largest magnitudes include $\tilde{o}_{ij}$, cc_{ij}^{S}, and I_{FM}. Weighted overlap is positively associated with tie strength, illustrating the greater the proportion of strength among overlapping friends of the focal nodes, the stronger the tie between the focal nodes, and showing evidence to support Granovetter's hypothesis. The sum of the clustering coefficients of the focal nodes is positively associated with tie strength, meaning tie strength decreases as the amount of clustering in the non-overlapping friendship circles increases. This provides quantitative evidence of Bott's hypothesis in a novel population. Finally, the predictor I_{FM} is negatively associated with tie strength, indicating that on average, female-male ties are weaker than male-male ties, which were used a reference group.

$$log(\mathbb{E}[w_{ij}]) = 1.62 + 2.41\tilde{o}_{ij} - 1.38cc_{ij}^{S} - 0.2\mathrm{I}_{\mathrm{FM}} \tag{3}$$

CDR Call Network. The CDR call network contained 2,276,495 nodes and 12,345,848 edges. Age was available for 89.25% of the individuals and had a mean of 48.2 (sd = 18.2) years. Of the 89.03% of individuals whose sex was recorded, 52.51% were male. Billing zip code was available for 99.35% of individuals. Overall, only 7.5% of nodal attributes were missing for this data set, and we therefore decided to perform imputation. Individuals in the CDR call network could have any combination of age, sex and billing zip code information missing. We used RF classification to impute sex and RF regression to impute age. Because of the abundance of billing zip code possibilities, rather than imputing billing zip code directly, we created a paired billing zip code dichotomous variable equal to 1 if the two focal nodes had the same billing zip code and 0 if they did not. We then used RF classification to impute paired billing zip code. After imputation, we sampled 500,000 of the 12,345,848 edges to be used as focal ties, excluding isolated ties, to limit computational expense. This resulted in a total of 496,941 ties. We then calculated the bow tie metrics using all of the nodes and ties contained in the bow tie structure surrounding each focal tie. Because the bow tie is a local structure, and none of the metrics used rely on global network topology, the topology of the network was not changed for the computations and subsequent analyses. Additionally, because we took a random sample of all edges in the network, the focal ties and associated bow ties used in the analyses are representative of the network as a whole. Similar to the India data set, three models were fit with RF regression both before and after nodal attribute imputation for each measure of tie strength and are denoted Models 1–3. Figure 3 shows the accuracy for RF regression after imputation for all three models and each measure of tie strength. The difference in accuracy for all models is very minimal and only one curve is visible for each tie strength measure. Within 0.05 units (a 5% difference between empirical and predicted tie strength), an accuracy of 61% was achieved for normalized tie strength, and 56.7% for averaged tie strength. Within 0.1 units, an accuracy of 76.5% was achieved for normalized tie strength and 77.3% for averaged tie strength. Accuracy for all models and both tie strength measurements before and after imputation are shown in Supplementary Figs S1 and S2. Imputation has a smaller impact on accuracy for this data set in all cases.

Feature importance for each of the three models after imputation is shown in Fig. 3. The black vertical line represents the value of importance if every predictor were equally informative. The most informative predictors in each model are s_{ij}^{S}, s_{ij}^{D}, n_{ij}^{S} and k_{ij}^{S}, with $\tilde{o}_{ij}$ and a_{ij}^{S} slightly more informative than the null importance value in models 1 and 3. This suggests focal node strength, degree and number of non-overlapping friends are the aspects of the bow tie most predictive of tie strength in this network. Feature importance plots for all models and all tie strength measures before and after imputation are presented in Supplementary Figs S1 and S2.

For each measure of tie strength, three different models, denoted Models A–C, were fit using linear regression methods following imputation. Model A denotes the full model that was fit using OLS regression. Model B was fit using LASSO and Model C using ridge regression. Because the distributions of normalized and averaged tie

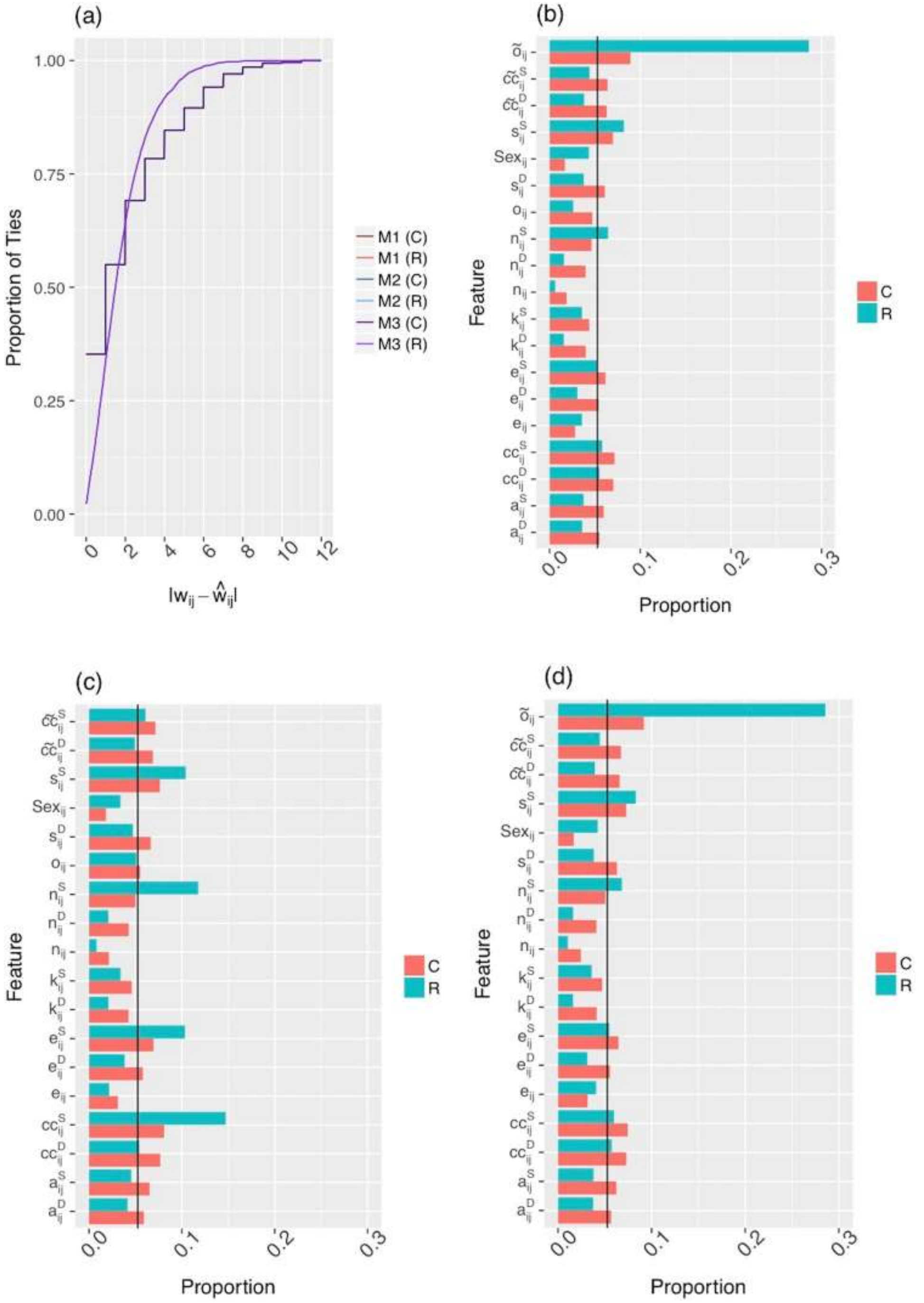

Figure 2. Accuracy and feature importance plots for the India social network. Accuracy, measured as the absolute difference between empirical tie strength (w_{ij}) and predicted tie strength ($\hat{w}_{ij}$), for Models 1–3 using both RF regression (R) and classification (C) after imputation is shown in (**a**). Feature importance using RF regression and classification after imputation are shown for Model 1 (**b**), Model 2 (**c**) and Model 3 (**d**). The horizontal bars represent how informative the predictor is with a longer bar meaning more informative. The black vertical line represents the value of an equilibrium or null importance if every predictor were equally informative.

strength are highly skewed for this data set, we first log-transformed each measure of tie strength and then centered them around the mean. All predictors were standardized (centered around the mean with unit variance) before fitting models B and C. Implementing LASSO and ridge regression require the selection of tuning parameters that determine the extent of shrinkage administered when calculating coefficient estimates. As the tuning parameter approaches 0, the corresponding coefficient estimates match the OLS estimates. In this extreme, the amount of bias is minimal, if nonexistent, but the amount of variance is comparatively high. As the tuning parameter is increased, the values of the coefficients decrease and approach 0 once the tuning parameter is sufficiently

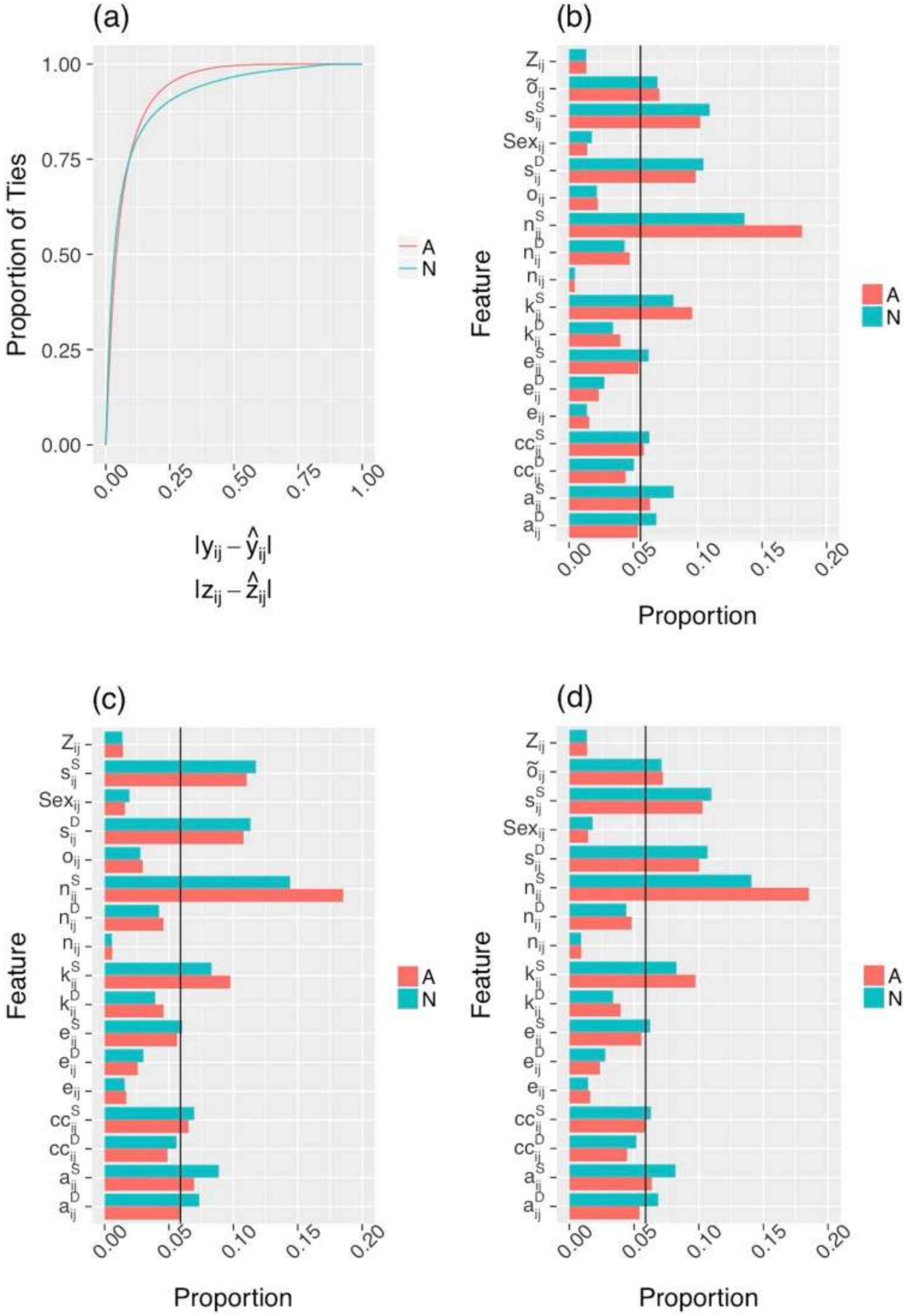

Figure 3. Accuracy and feature importance plots for the CDR call network with normalized (N) and averaged (A) tie strengths. Accuracy, measured as the absolute difference between empirical tie strength (y_{ij}, z_{ij}) and predicted tie strength ($\hat{y}_{ij}$, $\hat{z}_{ij}$), for all three models using RF regression after imputation is shown in (**a**). Note that only one curve is visible for each strength measure since the accuracy of all three models is indistinguishable. Feature importance using RF regression after imputation are shown for Model 1 (**b**), Model 2 (**c**) and Model 3 (**d**).

large. In this extreme, bias is increased but variance in the estimates is decreased. The optimal choice for a tuning parameter balances the amount of bias and variance and can be selected via cross-validation. We performed 10-fold cross validation to select values of the tuning parameters λ^L and λ^R. The values of the LASSO coefficients as a function of λ^L and, as a more interpretable measure, the l_1 penalty $\|\hat{\beta}_L\|/\|\hat{\beta}\|_1$ which represents the amount of shrinkage, are shown in Supplementary Figs S3 and S4. The values of the ridge regression coefficients as a function of λ^R and the l_2 penalty $\|\hat{\beta}_R\|/\|\hat{\beta}\|_2$ are shown in Supplementary Figs S3 and S4. Significant predictors, their coefficients, adjusted R^2 values and the values of the tuning parameters for models B and C are presented in Supplementary Table S1. Equations (4–6) show the fitted regression equations for normalized tie strength, y_{ij}, for OLS, LASSO and ridge regression respectively. Similarly, (7–9) show the fitted regression equations for averaged tie strength, z_{ij}, for OLS, LASSO and ridge regression respectively.

$$\mathbb{E}(y_{ij})_{OLS} = -0.35k_{ij}^{D} - 0.25s_{ij}^{S} + 0.29cc_{ij}^{D} + 0.23Z_{ij} + 0.27o_{ij} \tag{4}$$

$$\mathbb{E}(y_{ij})_{LASSO} = -0.33k_{ij}^{D} - 0.25s_{ij}^{S} + 0.23cc_{ij}^{D} + 0.23Z_{ij} + 0.21o_{ij} \tag{5}$$

$$\mathbb{E}(y_{ij})_{RIDGE} = -0.35k_{ij}^{D} - 0.25s_{ij}^{S} + 0.29cc_{ij}^{D} + 0.23Z_{ij} + 0.27o_{ij} \tag{6}$$

$$\mathbb{E}(z_{ij})_{OLS} = -0.35k_{ij}^{D} - 0.25s_{ij}^{S} + 0.29cc_{ij}^{D} + 0.23Z_{ij} + 0.27o_{ij} - 0.2s_{ij}^{D} \tag{7}$$

$$\mathbb{E}(z_{ij})_{LASSO} = -0.21k_{ij}^{D} - 0.39s_{ij}^{S} + 0.24cc_{ij}^{D} + 0.23Z_{ij} \tag{8}$$

$$\mathbb{E}(z_{ij})_{RIDGE} = -0.27k_{ij}^{D} - 0.49s_{ij}^{S} + 0.36cc_{ij}^{D} + 0.24Z_{ij} + 0.28o_{ij} + 0.31s_{ij}^{D} \tag{9}$$

For normalized tie strength, λ^R was sufficiently large such that no shrinkage was implemented, and the estimated ridge regression coefficients are equivalent to the OLS estimates. The amount of LASSO shrinkage was approximately 12%, resulting in slightly different coefficient estimates. In all models, o_{ij}, k_{ij}^{D}, s_{ij}^{S}, cc_{ij}^{D} and Z_{ij} were significantly associated with tie strength. Edge overlap is positively associated with tie strength in all models, showing that as the proportion of common friends two individuals share increases, so does the strength of the tie between the two individuals, supporting Granovetter's hypothesis. Tie strength is negatively associated with s_{ij}^{S} which suggests that as the focal nodes expand their social circles and the time spent interacting with friends, the weaker the tie between them; more evidence to support Bott's hypothesis. The positive association between Z_{ij} and tie strength implies having the same billing zip code increases the strength of a tie and could suggest a geographical impact on tie strength.

Here, cc_{ij}^{D} is positively associated with tie strength meaning the more dissimilar the non-overlapping clustering coefficients of the focal nodes, the stronger their tie. Lastly, the R^2 values for these models are on the lower side (0.112 on average). This could be due to the network being constructed with phone-based communication rather than face-to-face interactions among highly clustered villagers. Furthermore, quantifying tie strength for CDR data is currently still rather ambiguous; the operationalization of using communication as a proxy for tie strength has not yet been validated[20]. An alternate measure of tie strength may increase the R^2 values.

Discussion

In this work, we introduce the social bow tie; a novel framework we use to perform a comprehensive analysis of the association between network structure and tie strength. Our framework decomposes a social network into a collection of nodes and ties immediately surrounding each network tie. This utilization of local structure produces easily interpretable metrics that quantify social perspectives of tie strength and allows for analyses that are computationally feasible for networks of any size. Through machine learning and regression methods including LASSO and ridge regression, we determine which properties of the bow tie structure are the most predictive of tie strength in two different types of social networks; a contact network of Indian villagers and a nationwide call network of European mobile phone users.

Overall, both data sets provide evidence to support the weak ties hypothesis and the Bott hypothesis. Following Granovetter, we find that the more friends two individuals share, the stronger their tie. Following Bott, the more tightly-knit their individual social circles, the weaker their tie. In addition, we find that the bow tie framework provides metrics that predict tie strength with high accuracy for both networks.

In future work, it would be interesting to apply the bow tie framework to a social network of married couples. In this case the dominant strong tie has properties that are not seen in more casual social ties, namely the individuals constitute a particularly strongly defined social institution that has both emotional (romantic attachment) as well as structural (e.g. common responsibility for children and common ownership of capital investments such as a home) elements that provide it resiliency. This would enable testing of the original version of Bott's hypothesis, rather than a generalized form as we present here. It would also be interesting to test if the strength of in-person ties behaves similarly for the mobile phone call network.

References

1. Dunbar, R. I. M. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution* **22**, 469–493 (1992).
2. Li, N., Feng, X., Ji, S. & Xu, K. *Modeling Relationship Strength for Link Prediction*, 62–74 (2013).
3. Linyuan, L. & Tao, Z. Link prediction in weighted networks: The role of weak ties. *EPL Europhysics Letters* **89**, 18001 (2010).
4. Sá, H. R. d. & Prudêncio, R. B. C. Supervised link prediction in weighted networks. In *paper presented at the 2011 International Joint Conference on Neural Networks, San Jose, CA* (New York, NY: IEEE, 2011, October).
5. Granovetter, M. The strength of weak ties. *American Journal of Sociology* **78**, 1360–1380 (1973).
6. Marsden, P. V. & Campbell, K. E. Measuring tie strength. *Social Forces* **63**, 482–501 (1984).
7. Wellman, B. & Wortley, S. Different strokes from different folks: Community ties and social support. *American Journal of Sociology* **96**, 558–588 (1990).

8. Ellison, N. B., Steinfield, C. & Lampe, C. The benefits of facebook 'friends': Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* **12**, 1143–1168 (2007).
9. Lin, N., Vaughn, J. & Ensel, W. Social resources and occupational status attainment. *Social Forces* **59**, 1163–1181 (1981).
10. Xiang, R., Neville, J. & Rogati, M. Predicting tie strength in a new medium. In *CSCW '12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, Seattle, WA* (New York, NY, ACM, 2012, February).
11. He, J., Chu, W. W. & Liu, Z. Inferring privacy information from social networks. In *Intelligence and Security Informatics, Lecture Notes in Computer Science*, vol. 3975, 154–165 (Springer, Berlin, Heidelberg, 2006).
12. Bott, E. Family and Social Network: Roles, Norms and External Relationships in Ordinary Urban Families (Abingdon: Routledge, 1957).
13. Rogler, L. & Procidano, M. The effect of social networks on marital roles: A test of the bott hypothesis in an intergenerational context. *Journal of Marriage and the Family* **48**, 693–701 (1986).
14. Udry, J. R. & Hall, M. Marital role segregation and social networks in middle-class middle-aged couples. *Journal of Marriage and Family* **27**, 392–395 (1965).
15. Schonpflug, R. K. S., Ute & Schulz, J. Perceived decision-making influence in turkish migrant workers and german workers families:the impact of social support. *Journal of Cross-Cultural Psychology* **21**, 261–282 (1990).
16. Pappalardo, L., Rossetti, G. & Pedreschi, D. How well do we know each other? detecting tie strength in multidimensional social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM 12, 1040–1045 (IEEE Computer Society, Washington, DC, USA, 2012).
17. Kahanda, I. & Neville, J. Using transactional information to predict link strength in online social networks. In *Proceedings of the Third International ICWSM Conference*, 74–81 (Association for the Advancement of Artificial Intelligence, 2009, June).
18. Luarn, P. & Chiu, Y.-P. Key variables to predict tie strength on social network sites. *Internet Research* **25**, 218–238 (2015).
19. Gilbert, E. Predicting tie strength in a new medium. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW 12, 1047–1056 (ACM, New York, NY, USA, 2012).
20. Wiese, J., Min, J.-K., Hong, J. & Zimmerman, J. *Assessing Call and SMS Logs as an Indication of Tie Strength* (2014). Unpublished manuscript.
21. Banerjee, A., Chandrasekhar, A., Duflo, E. & Jackson, M. The diffusion of microfinance. *Science* **341** (2013).
22. Broder, A. *et al*. Graph structure in the web. *Comput. Netw.* **33**, 309–320 (2000).
23. Onnela, J.-P. *et al*. Structure and tie strengths in mobile communication networks. *PNAS* **104**, 7332–7336 (2007).
24. Mattie, H. & Onnela, J.-P. *Edge Overlap in Weighted and Directed Social Networks* (2017). Unpublished manuscript.
25. Saramaki, J., M., Kivela, J.-P. O., Kaski, K. & Kertesz, J. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E* **75**, 027105-1–027105-4 (2007).
26. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
27. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer New York Inc., New York, NY, 2001).

Acknowledgements

The authors thank Telenor for making available anonymized data for this study, and Banerjee *et al*. for making the India data set publicly available. H.M. was supported by the National Institute of Health, project number 5T32AI007358-28.

Author Contributions

J.-P.O. supervised all work; K.E.-M. processed raw mobile phone data. H.M., K.E.-M., R.L. and J.-P.O. performed research; H.M. and J.-P.O. analyzed data; H.M., K.E.-M., R.L. and J.-P.O. wrote the paper. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-27290-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.