Urena Carrion, Javier; Saramäki, Jari; Kivelä, Mikko

Estimating tie strength in social networks using temporal communication data

Published in:
EPJ Data Science

DOI:
10.1140/epjds/s13688-020-00256-5

Published: 03/12/2020

Document Version
Publisher's PDF, also known as Version of record

Ureña-Carrion et al. *EPJ Data Science* (2020) 9:37
https://doi.org/10.1140/epjds/s13688-020-00256-5

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Estimating tie strength in social networks using temporal communication data

Javier Ureña-Carrion[1*], Jari Saramäki[1] and Mikko Kivelä[1]

*Correspondence:
javier.urenacarrion@aalto.fi
[1]School of Science, Aalto University, Otakaari, Espoo, Finland

Abstract

Even though the concept of tie strength is central in social network analysis, it is difficult to quantify how strong social ties are. One typical way of estimating tie strength in data-driven studies has been to simply count the total number or duration of contacts between two people. This, however, disregards many features that can be extracted from the rich data sets used for social network reconstruction. Here, we focus on contact data with temporal information. We systematically study how features of the contact time series are related to topological features usually associated with tie strength. We focus on a large mobile-phone dataset and measure a number of properties of the contact time series for each tie and use these to predict the so-called neighbourhood overlap, a feature related to strong ties in the sociological literature. We observe a strong relationship between temporal features and the neighbourhood overlap, with many features outperforming simple contact counts. Features that stand out include the number of days with calls, number of bursty cascades, typical times of contacts, and temporal stability. These are also seen to correlate with the overlap in diverse smaller communication datasets studied for reference. Taken together, our results suggest that such temporal features could be useful for inferring social network structure from communication data.

Keywords: Social networks; Tie Strength; Call Detail Records; Communication networks

1 Introduction

During the past few decades, the use of auto-recorded data, such as mobile phone logs or data from online platforms, has expanded our understanding of human dynamics and networks [1–4]. Such data have also been useful in applications ranging from spreading dynamics [5] to human mobility [6], recovery in disaster areas [7], and health-care optimization [8]. In particular in social network studies, the *strength of a tie* is a central concept associated with the qualitative value that people place on relationships. Tie strength is not, however, something that can be directly measured or quantified [9–11]. Therefore, one has to rely on proxies. For networks reconstructed from data on communication events, such as call networks, a common approach is to use a measure of *communication intensity* as a proxy [1, 3, 12–18]. Communication intensity can be defined as the total number of communication events or the total time spent communicating across a tie. One motiva-

Springer

tion behind this choice is that intense communication implies temporal and sometimes even financial commitment to a relationship [1]. Communication intensity is, however, an aggregated measure that discards a lot of possibly relevant information contained in the underlying time series of dyadic interactions. This temporal information is at the focus of the present paper.

In addition to such internal details of a social tie, the network structure surrounding a tie is known to be informative about the nature of the tie. The seminal paper "The strength of weak ties" [19] by Mark Granovetter was one of the first efforts to relate tie strength with local network structure and to connect the micro and macro levels by considering the role of weak ties in diffusion and social mobility. Granovetter argued that strong ties tend to be associated with overlapping circles of friendship, while weak ties serve as bridges between such circles. This implies that weak ties serve are more important for network-wide information diffusion than strong ties.

In this study, we assume that *the strength of a tie is a latent variable expressed both in patterns of dyadic interactions and in network topology*, the latter following Granovetter's overlap hypothesis [19]. This way, we use neighbourhood overlap as a benchmark that allows us to compare different characteristic features of communication events taking place on a tie: we use a feature's predictive capacity for neighbourhood overlap as proof of association with the latent strength of a tie. This approach is shown schematically in Fig. 1. Our method builds heavily from existing literature, yet the contribution of our work is to use a purely data-analytic approach, combining information about communication and network topology thorough the lens of Granovetter's theory.

This paper is structured as follows: first, we discuss how tie strength has been conceptualized and measured in previous research, both from the sociological and network-scientific perspectives. Then, we explore different modelling approaches to human communication which serve as a theoretical basis for our predictive features. We address the temporality of our data (a) as time series or sequences of interactions and (b) as events that occur within natural daily rhythms and weekly social cycles. Following this, we present results obtained for linking temporal features with neighbourhood overlap in a large-scale communication network. We determine the importance of different variables as proxies for topological tie strengths, and show that many alternative features that can be used in network construction as opposed to using a default number of contacts. We further test our hypothesis with additional communication datasets. We conclude with discussion.

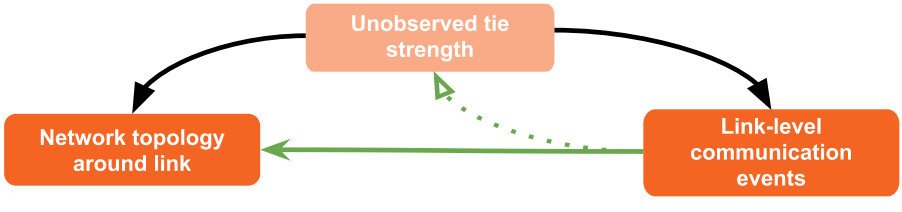

Figure 1 Representation of our conceptualization of tie strength as a latent variable that drives both network topology and patterns of human communication (black arrows). While the tie strength is unobserved, we argue that using characteristic temporal features of human communication to predict topology (green solid arrow) allows us to determine those features that best reflect this latent variable (green dashed arrow)

2 Measuring tie strength

2.1 Historical background

Despite its relatively intuitive definition in terms of emotional closeness [10], the strength of a tie is a sociological concept with no direct indicator and its measurement requires prior theoretical definition and empirical validation [9–11, 20]. We can broadly distinguish two methodological variants: early studies often borrow from social psychology and rely on self-reported surveys, whereas more recent studies build on the availability of large sets of auto-recorded and behavioural data, which have spawned a wide array of methods and lines of research.

The first conceptualizations of tie strength focused on intrinsic tie-level characteristics, such as relationship 'closeness' or kinship (e.g., relatives have strong ties while neighbors have weak) [10]. Alternatively, some researchers highlighted the effect of ties on the nodes, such as the provision of emotional support [21], or the ability to handle multiple contexts [19, 22]. In his paper, Granovetter did not delve deeply into the definition of tie strength, characterizing it as a "possibly linear" combination of four constituting dimensions—time, emotional intensity, intimacy, and reciprocity [19]. Many early studies analyzed social ties via standardized surveys that enquired about friendship, emotional support, frequency of contacts, or advise seeking [9, 10, 20, 23], while acknowledging the limitations of self-reported and, to a large degree, unilateral data for dyadic interactions [10, 23, 24]. Marsden and Campbell [10] used survey data to determine which proxies for Granovetter's dimensions were most strongly associated with self-identified tie strength, suggesting that tie strength could be a multidimensional concept.

Other lines of research have highlighted the temporal and dynamic aspect of human relationships, characterizing qualitative differences by relationship stages: initiation, maintenance and decay [23, 25, 26]. Gradual stages of reciprocity were identified as a key component in friendship formation [23, 27]. Burt [28, 29] argued that factors associated with strong ties (homophily, social status, embedding, and inertia) are also associated with slower tie decay, but that tie decay is guided by nodes via selection processes and learning of social routines. Notably, both relationship initiation and decay were conceptualized as involving topological changes in social networks [25]. Burt [28] found evidence that embedded ties were associated with slow decay, but that disruptions in embedding implied even faster decay. Some of these topological changes around tie decay were later be examined in dynamical contexts [17, 30, 31]. Moreover, even ties that were neither nascent nor decaying were established to be highly dynamic [26], with Wilmot arguing that relational stability does not imply that relationships are static, but that there is a minimal agreement about the relationship which is reflected on communication patterns [25, 32].

From a socio-psychological perspective, Feld [33] focused on how ties appear in social contexts that facilitate interaction, named *foci*. Tie strength was thus theorized to be determined by sociological roles: a small and constraining focus (such as a nuclear family) might imply higher strength, but the interaction of multiple foci explains the multiplexity of ties, thus conveying the idea that two people interact in different contexts and social groups. The concept of foci was more recently exploited by [30] to identify romantic partners, finding that people in romantic relationships have a focalized network structure—they both share a large number of common friends, but these friends belong to different foci, so they are *dispersed*, or not connected among themselves.

In the recent decades, technological advances and the advent of telecommunication devices and social media platforms have provided access to an unprecedented collection of auto-recorded data [34]. This has generated new methodologies and conceptualizations of tie strength, which depend not only the underlying social network, but also on the data source with the appearance of distinct communication channels, such as phone calls or emails, but also of specific social platforms, such as Facebook, MySpace or Twitter. This has opened various previously unattainable lines of research, such as research on large-scale network properties [1], characteristics of human communication networks [35], temporal networks [36], link prediction [37], and link decay [38]. In these cases, many studies adopted quantifications of tie strength in terms of communication intensity, such as the total number or duration of contacts [1, 20, 31, 36, 39]. Other approaches have complemented the use of auto-recorded data with either surveys on emotional closeness [18, 20, 40, 41] or tagged human interactions in online platforms, such as interactions with spouse or close friend [30, 34], while other studies have determined features inspired by Granovetter's four dimensions; Navarro [31], for instance, determined that strong ties were those that were unlikely to decay and identified features that predicted this.

2.2 Tie strength and network topology

In this paper, we focus on untagged human interactions, where our goal is to infer the latent tie strength from behavioural features of communication. We conceptualize *tie strength* as a latent variable that manifests independently in both network structure and communication patterns, so that strong ties are embedded in dense network communities while weak ties serve as inter-community bridges [19, 22]. The network structure and communication patterns are considered independent in the sense that no data used for computing network structure around the tie is used again for computing the temporal features of the tie's communication patterns. Under our framework, the embeddedness or friendship overlap of a tie serves as a baseline that relates tie strength to features of communication. In this sense, variables with good predictive performance of topological features serve as better proxies for tie strength, at least to the extent they are reflected in local network topology.

We measure embeddedness using topological overlap [1], O_{ij}, which is defined as the Jaccard similarity of the sets of neighbors of two nodes i and j, a measure that can be interpreted as the percentage of common neighbors around a tie. Formally the topological overlap is defined as

$$O_{ij} = \frac{|\mathcal{N}_i \cap \mathcal{N}_j|}{|\mathcal{N}_i \cup \mathcal{N}_j|}, \tag{1}$$

where $\mathcal{N}_i$ is the set of neighbors of node i. The Granovetter effect—the increase in embeddedness along with tie strength—was previously observed using overlap and number of calls (w_{ij}) and total call time (l) as proxies for link weights [1, 39]. Previous research has also found that different communication patterns entail topological changes in social networks [25, 30, 31, 39], and indeed tie evolution has also been associated with distinct features of human communication for both tie creation and decay [31].

On this matter, our focus is not on detecting topological change. This is because (i) topological variations have been shown to occur over long periods of time [30, 31, 39] which correspondingly requires long-term longitudinal data and (ii) they entail the additional

problem of uncoupling bursty communication patterns from changes in the underlying social structure [31, 35]. This issue is heightened by different social strategies empirically observed in communication networks, where *explorers* display a large turnover of weaker ties, while *keepers* prefer a smaller circle of stronger ties [39]. In addition, we know overlap to be a decreasing function of the aggregation window for communication networks [13]. To address these problems, we assume that tie strength remains constant during our observation period, which we expect to be true for most relationships in a span of a few months [25, 26, 32], and provide a dynamic measure of overlap that penalizes ties that are not active over a long period. We measure overlap in a dynamic manner by establishing a smaller aggregation window, ΔT, which we shift over the full period and to obtain a time series of overlap values $\{O_{ij}^t\}_{t=1}^{N_t}$. We use the average of our time series as a measure of temporally averaged overlap, $\hat{O}_{ij}^t$. This variable emphasizes edges that are relatively close in time. We obtained dynamic overlap with $\Delta T = 1$ month, which we justify since empirical evidence on similar datasets [13] has found overlap to become relatively stable at an aggregation window of this size. To serve as a baseline, we repeated the same experiments using the static overlap over the full observation on Additional file 1.

2.3 Source data

Our main data set is a single large-scale database of Call Detail Records, from which we derive the key results of this paper. For comparison, we also use public or previously-published communication datasets that, although smaller in scope. We selected the latter datasets by systematically examining network repositories and other available sources. We excluded datasets that did not meet validity criteria, namely, that the data contained sufficient topological information around ties, and that the data displayed evidence of the Granovetter effect when using contact counts for tie strengths — an initial assumption of our analysis. In more detail:

1 We use a Call Detail Records (CDRs) database from a single operator in a European country [1], with an observation period of four months during 2007 and a market share of 20%. CRDs are communication logs recorded by mobile service providers, where basic information of the interaction is sequentially stored, including, e.g., caller, callee, timestamp and duration. CDRs from single operators are functionally a statistical sample of a complete dataset of interactions [1]. Despite lacking a full network, our dataset does provide full ego networks centered on our operator's subscribers. We thus focus our study on the strength of ties that fully belong to our operator (both nodes in a tie are subscribers), involving ~ 6.5 million nodes and ~ 26.4 million ties; however, for network topology we also use ties of non-company users, which correspond to an additional ~ 76 million nodes and ~ 530 million ties. This methodological choice guarantees that there is no bias related only single operator links being included in the overlap calculation. This mitigates the concerns that our dataset is not a random sample - because family ties, friendship recommendations and regional differences in market share may be drivers when customers choose a mobile service provider, and these differences might result in biased estimates of overlap.

2 We use four additional datasets, which have all been anonymized:
   - *Forum*: The Forum dataset [42] contains activity of 6269 users in an online community during a period of around 7 years. The data is anonymized as to

only contain interaction information (user ID, time since beginning of collection period). Users in the community have access to profile pages that includes movie tastes, list of friends and rated movies. A contact record is created if user i comments on a thread posted by user j. For this case we found strong evidence for the Granovetter effect.

- *Enron*: Enron dataset obtained from the Network Repository [43], an email dataset of communication within a company. The dataset contains 20165 users and 297456 edges. Note, however, that the sampling of the data might give more importance to certain individuals within the company, so we lack knowledge to know how this might affect the end results. This dataset displayed mild evidence for the Granovetter effect.
- *FB*: A dataset of wall-posting on Facebook in New Orleans [44]. The network contains 40981 nodes and 183412 edges, and a link is created if a user has posted on another user's wall. The dataset was obtained by crawling public profiles in January 2009 and focuses on activity during each link's first year of communication. This dataset showed strong evidence for the Granovetter effect.
- *Copenhagen*: Copenhagen Networks Study [45], a dataset of the communication of more than 700 university students during four weeks, that includes text messages, phone calls, Facebook friendships and proximity data. We use text messages as our communication data, and we construct a network using a combination of text messages, phone calls and Facebook friendships, where at least one message, one call or Facebook friendship suffices for the creation of an edge. For this dataset we found weak evidence for the Granovetter effect, possibly due to the small sample of students.

We focus on the mobile CDR database for studying communication patterns and topology for three main reasons. First, the sampling is independent of social foci [33]; it is thus more representative of many everyday social interactions, and not biased by schools, workplaces or communities. Second, our CDR dataset potentially captures the main communication channel at the time of collection [36]. Taking place in 2007, we expect a smaller effect of multiple communication channels, which have become more common in recent years [36, 45, 46]. Last, the large-scale of the dataset allows us to extract large amounts of information (for example about weekly behaviour), which might not be available otherwise. For these reasons, we expect our CDR data to be a suitable tool for understanding communication patterns and network topology, and the Granovetter theory to hold.

The four additional datasets allow us to expand the scope of our analysis and test whether similar relationships hold under different circumstances, including much narrower social foci, different communication channels and both online and offline social networks. While one cannot generally expect the Granovetter picture to be valid for networks that are not necessarily good proxies for "everyday" social interactions, the four chosen datasets were nevertheless in line with this picture.

3 Features of human communication

Our aim is to determine features that might encode information on the tie strength not captured by intensity variables. Figure 2 illustrates this idea by showcasing ties in our data of similar communication intensity w that differ both in overlap and communication patterns. In the following, we expand on these temporal features of human communication,

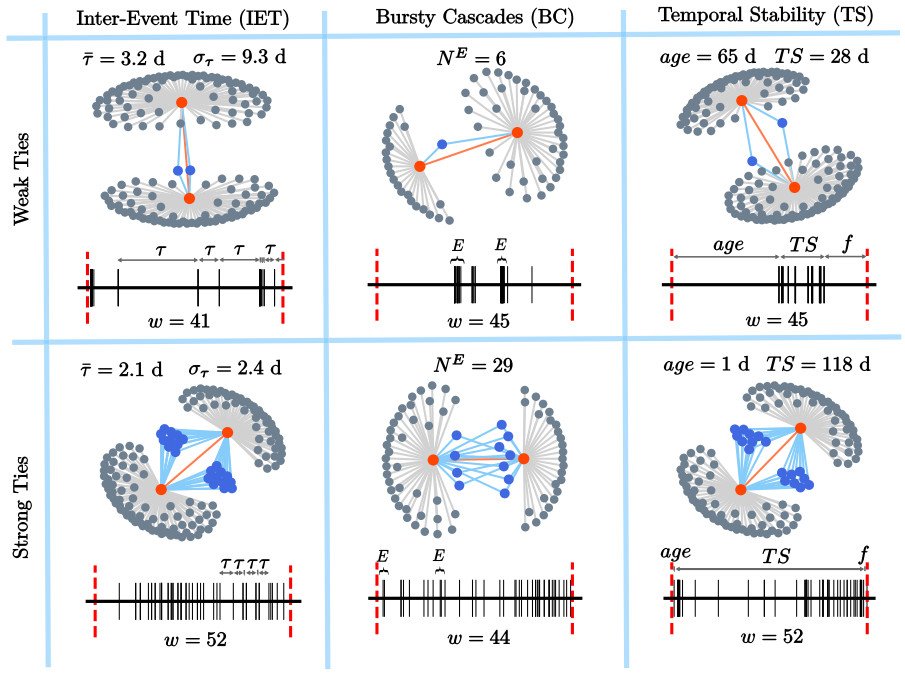

Figure 2 Differences in neighbourhood overlap (*top* and *bottom*) and selected temporal features of human communication (columns) for ties of similar communication intensity ($w \in [40, 55)$). Weak (*top*) and strong (*bottom*) ties are defined by low and high overlap, respectively, where we analyze communication patterns of red ties (ij), blue ties represent common neighbors ($\mathcal{N}_i \cap \mathcal{N}_j$), and grey ties represent neighbors of either i or j but not both. (*Left*) IET distribution; we model the time between two consecutive calls as a homogeneous process via a random variable $\boldsymbol{\tau}$, and obtain statistics on the IET distribution, such as $\bar{\tau}$ and σ_τ, both measured in days. (*Center*) Bursty Cascades; given a parameter Δt, we identify event *bursts* E as sub-sequences of calls that are placed at most Δt seconds within each other. N^E is the number of events. (*Right*) Temporal stability; we focus on the first and last events and their distribution within the observation period (red dashed lines), and determine the *age* as the time until the first call, the temporal stability *TS* as the time window where we observed events, and freshness f as the time between the last event and the end of the observation period

and use them as predictors of overlap, comparing them with the widely-used number of contacts as a communication-intensity measure.

A key assumption of this work is that differences in the strengths of ties are reflected in communication patterns of dyadic interactions. Based on these data, we collected variables from existing literature that model different aspects of human communication and developed some new indicators. We roughly divide our approach in two: measures building on the sequential nature of our data and measures focusing on daily and weekly behaviour.

3.1 Intensity features of human communication

Features related to communication intensity have commonly been used as a proxy for tie strength [1, 3, 13]. We denote the number of contacts as w, as this is commonly used as link weight in social network analysis. The definition of the number of contacts will depend on the communication type of communication activity, e.g., calls for the CDR network. We further analyze communication intensity in terms of total call time $l = \sum_{i=1}^{w} l_i$ where call i has length l_i, as well as average call time $\hat{l} = l/w$.

We also characterize the reciprocity r as an intensity feature [27, 31], which we measure via

$$r_{ij} = \left| \frac{\vec{w}_{ij}}{w_{ij}} - \frac{1}{2} \right|, \tag{2}$$

where $\vec{w}_{ij}$ is the number of calls placed by i to j, so that $r \approx 0$ implies that both users placed a similar amount of calls, while $r \approx 0.5$ reflects an imbalance.

3.2 Sequential features of human communication

At the level of ties, CDRs and our other data sources record a time series of events. Most of the measures based on these time series are based on the intuition that regular contacts are more significant than for example brief periods of large contact intensity; we exemplify some of these modelling approaches in Fig. 2. In this section, we first focus on measuring the number of time periods during which the tie has been active. Second, we consider the time elapsed between consecutive contacts via the inter-event time (IET) distribution. Third, we focus on correlated bursty behaviour and memory effects, using the distribution of event *bursts*. Last, we focus on behavioural changes within the observation window, with variables that have been previously associated with tie creation and decay.

3.2.1 Counting active periods

The regularity of a time series can be measured by counting the active periods, such as hours or days, with at least one contact. We record the number of hours and days with events, a_h and a_d, respectively. Since we know human communication to be bursty [5, 47–49], this aggregating process serves to remove temporal correlations to different degrees. These variables also allow for the incorporation of different communication channels, such as phone calls and text messages [46].

3.2.2 Inter-event time distribution

We measure the IET, the elapsed time between consecutive calls (τ), depicted in Fig. 2 (*left*). Given the set of interaction times $\{t_0, t_1, \ldots, t_n\}$, we obtain the kth inter-event time $\tau_k = t_{k+1} - t_k$, and in practice may estimate moments from this distribution from the empirical observations $\{\tau_k\}$.

The IET distribution encodes uncorrelated information about the times between consecutive calls. This uncouples temporal correlations between events [50] while discarding possible memory effects between consecutive inter-call times. This allows us to obtain general call patterns such as the mean IET $\bar{\tau}$ and the standard deviation of the IET σ_τ, where a small $\bar{\tau}$ would imply more frequent communication, which has been theorized to occur when ties are strong [25]. Previous research has estimated the IET distribution to be heavy tailed [48, 51] and bursty, so that short spikes of activity are followed by long periods of inactivity [47, 48]. In this sense, the IET distribution provides a natural way to characterize uncorrelated burstiness via the burstiness coefficient $B = \frac{\sigma_\tau - \bar{\tau}}{\sigma_\tau + \bar{\tau}}$ [48], which takes value $B = -1$ for completely regular IETs, $B = 0$ for Poissonian behaviour, and $B = 1$ for completely bursty or irregular behaviour. A related measure is the average relay or waiting time τ_R, which is defined as the time between a random point in time and the next event. It can be used as a local measure of the speed of information spreading over the link, and when normalised with $\bar{\tau}$ it has been shown to be a non-linear function of B [52].

3.2.3 Bursty cascades

Temporal correlations, neglected by the IET distribution, are common in human communication [47]. Our next set of features places a larger importance on bursts as determined via a parameter Δt. Karsai *et al.* [47] define a bursty cascade by the number of consecutive communication events E that took place within a time period of Δt or less; in other words, events k and $k+1$ are part of the same bursty train iff $\tau_k = t_{k+1} - t_k \leq \Delta t$, as depicted in Fig. 2 (*b*).

This approach has been used to find that $P(E)$, the distribution of the number of events in a bursty cascade, is also heavy-tailed over a range of Δt values [47, 50]. In contrast, if the event times are uncorrelated but follow the same IET distribution, there is an exponential decay for $P(E)$. The structure of correlations that can be constructed from the bursty cascades at different resolutions Δt is completely independent of the IET distribution [50]. This allows for a flexible characterization of human communication, where the main focus is not on calls, but on call cascades. In this respect, this shift of focus provides new features of communication frequency via the number of cascades, but also via how calls are distributed within cascades.

We use a set of variables related to bursty cascades, including the mean number of events per cascade $\bar{E}$, the standard deviation σ^E, the coefficient of variation $CV^E = \frac{\sigma^E}{\bar{E}}$ and the number of bursty cascades N^E. We chose to use $\Delta t = 26$ hours, since preliminary tests showed that this yielded the best association with overlap. These results, available in the Additional file 1 (SI), corroborate that $P(E)$ is not overtly sensitive to the choice of Δt.

3.2.4 Temporal stability

The above approaches implicitly assume that behaviour doesn't change in time, that is, they measure communication activity while assuming that the underlying social relationship remains constant. As previously stated, it is not trivial to disentangle bursty communication patterns from the underlying dynamic relationship, where long IETs might be interpreted as tie decay [35]. We may, however, measure behavioural changes during the observation window, for which we use two sets of variables. For the first set of variables, we divide the observation window into three sub-intervals, measuring a) the *age* of a tie as the first observed communication event [53] b) the temporal stability (TS) of a tie as the elapsed time between the first and last communication events, and c) the freshness of a tie f as the time elapsed between the last communication event and the end of the observation window [31, 38, 41, 53]. For the last variable, we use relative freshness $f^r = f/\bar{\tau}$, which allows us to compare the time elapsed with no communication with the average IET, a metric which has been used to predict tie decay [31].

3.2.5 Distribution of bursty cascades

Next, our goal is to characterize *when* communication takes place within the observation window, in a similar fashion to temporal stability features. The previous measures, however, used only the first and last communication events, while we will now work on the whole set of interactions. We decouple correlated bursty behaviour by focusing on the distribution of bursty cascades within the observation period, as opposed to the distribution of calls.

We define our variables as follows: given a parameter Δt and a sequence of interaction timestamps $\{t_j\}_{j=1}^{w}$, where each t_j has been normalized to the interval $[0, 1]$ defined

by the observation window, we obtain a sequence of timestamps for bursty trains $\{t_i^*\}_{i=1}^{N^E}$, where t_i^* corresponds to the first observed event within bursty train i. We define the average interaction time $\bar{t} = \frac{1}{N^E}\sum_i t_i^*$, and the associated standard deviation σ_t. We found that overlap decreases for average interaction times that were skewed on the observation window (average values $\bar{t}_{ij}$ far from $t = 0.5$). For this reason, we included a feature that measures deviation from $t = 0.5$ as a test statistic for difference of means with unknown variance $T = \frac{\bar{t}-0.5}{\sigma_t\sqrt{N^E}}$. We use $\log(T)$ to penalize outliers.

3.3 Daily and weekly features

Human behaviour is regulated by the interplay of natural and social factors that determine different degrees of activity during, e.g., the day-night cycle or weekday-weekend cycle [54–57]. Our goal in this section is to determine whether these fluctuations are also reflected in network topology. We focus on two main sets of variables: first, we analyze differences in daily activity patterns, and second, differences in call profiles during the week.

3.3.1 Differences in daily patterns

Although humans typically follow 24-hour cycles determined by daylight, behaviour during these cycles has been found to be highly heterogeneous [58, 59]. In particular, there are prominent individual differences among the morningness or eveningness of people [55, 60, 61]; that is, the propensity to be more active during the morning or evening. We look for differences in daily call patterns of people forming dyads, and use these as a candidate measure for predicting tie strength. This variable is conceptually different from the previous ones as it is defined using information from two nodes instead of a single tie. Our hypothesis is that there are several reasons why people linked by strong ties have more similar daily call rhythms: people might have habitual calling patterns, the activities of friends might be synchronized through joint activities, or there might be latent drivers of call behaviour that are also associated with homophily [62, 63], such as age.

For each person, we compute a 24-hour daily distribution $P = (p^0, \ldots, p^{23})$ of the fraction of outgoing calls placed during each hour. For each tie, we then measure differences in the daily distributions by using the Jensen-Shannon Divergence (JSD), chosen for its ability to handle zero-valued probabilities. The JSD is defined for two discrete probability distributions P_0 and P_1 as

$$\mathrm{JSD}(P_0, P_1) = H\left(\frac{1}{2}P_0 + \frac{1}{2}P_1\right) - \frac{1}{2}\big(H(P_0) + H(P_1)\big), \tag{3}$$

where H is the Shannon entropy, $H(P) = -\sum_t p(t)\log(p(t))$.

3.3.2 Weekly activity profiles

Our last focus is weekly behaviour, where we identify times during the weekly cycle where a distinct call profile might be associated to higher/lower topological overlap. The motivation is that ties within different groups or foci might be associated with different call-placing patters: activity between colleagues can be expected to differ from that between family members or friends [33]. We follow a two-step procedure where we first divide the week into $7 \times 24 = 168$ hourly bins, and to each bin we assign the fraction of calls placed by both nodes in a tie. Unlike for the daily patterns, the focus is therefore on ties instead

of node-level behaviour. This high-granularity approach yields features that are too sparse to be interpretable; for this reason, as a second step we perform dimensionality reduction based on the overall call profiles of the whole dataset. We base this dimensionality reduction on our 168-feature correlation matrix and their association with overlap. For details, see SI.

4 Results

4.1 Clustering of weekly call patterns

Figure 3 depicts our results on how different weekly call profiles are associated with different overlap values. After our dimensionality reduction process, we obtained 15 clusters $\{C_i\}_{i=1}^{15}$ which constitute a weekly call profile vector C^* for each tie; we normalize the component contributions so that $|C_{ij}^*| = 1$. We find that there is heterogeneity in the association between the call profiles and overlap: the fraction of a tie's calls that belong to cluster C_{12} (weekend late morning and early afternoon) correlates positively with the overlap, whereas there is a low negative correlation for late-night calls (cluster C_1).

4.2 Predicting overlap from tie features

Our goal is to predict topological overlap using features computed for ties, and to compare their performance to simple communication intensity measures. Table 1 contains a list of the features used in our study, while Fig. 4 (i) depicts the Pearson's correlation coefficient for our features. We find high degrees of correlation for certain groups of variables — such as a_d and a_h with w, which is expected—, yet their association to overlap differs. To show that such features have explanatory power beyond the number of contacts w, we have stratified ties into groups based on w and studied how the overlap depends on the variable associated with each feature within the groups. This dependence is shown for

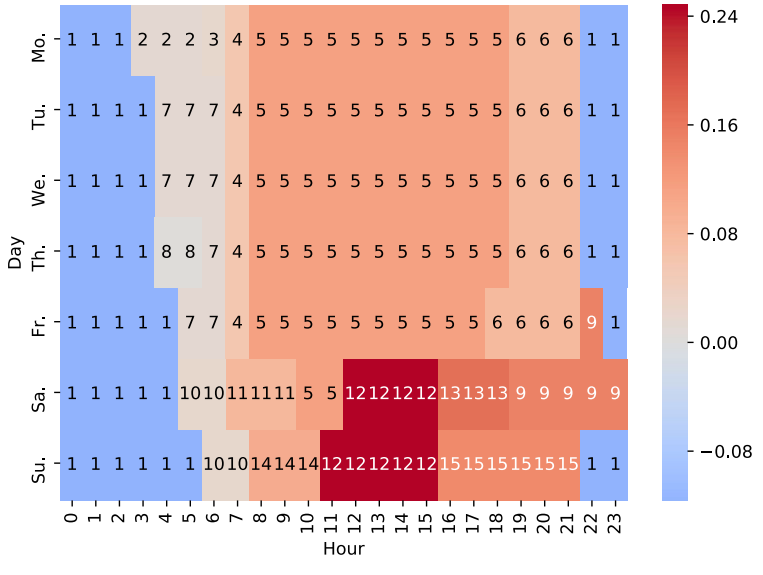

Figure 3 Composition of weekly call profiles of social ties and their association with the neighbourhood overlap. Each bin represents an hour of the week (x-axis: hours, y-axis: days), and the number inside the bin is the corresponding cluster index. The bin's color indicates the level of Pearson's correlation of each tie's fraction of events in the bin's cluster with the overlap. *E.g.*, a tie's topological overlap correlates positively with the fraction of calls across that tie that take place between noon and 4 PM on weekends

Table 1 Features of human communication used in our analysis. Our feature types - Intensity (I), Active Periods (AP), Inter-event time (IET), Temporal Stability (TS), Bursty Cascades (BC), Distribution of bursty cascades (DBC), differences in daily patterns (DP) and clusters for weekly activity

Type	Variable	Name	Cluster	Description
I	w	Number of calls/contacts	C_1	Late night and early morning
I	l	Total call duration	C_2	Monday early morning
I	$\bar{l}$	Average call duration	C_3	Monday early morning
I	r	Reciprocity	C_4	Weekday 7 am
AP	a_d	Active days	C_5	Weekday afternoon
AP	a_h	Active hours	C_6	Weekday evening
IET	$\bar{\tau}$	Mean IET	C_7	Weekday early morning
IET	σ_τ	Std. Dev. of IET	C_8	Thursday early morning
IET	B	Burstiness Coefficient	C_9	Weekend evening
IET	$\bar{\tau}_R$	Average Relay Time	C_{10}	Weekend morning
TS	$\hat{f}$	Relative freshness	C_{11}	Saturday Morning
TS	age	Age	C_{12}	Weekend afternoon
TS	TS	Temporal Stability	C_{13}	Saturday late afternoon
BC	N^E	Number of busty events	C_{14}	Sunday morning
BC	$\bar{E}$	Average calls per bursty event	C_{15}	Sunday afternoon
BC	σ^E	Std. Dev. of event distribution	C^*	Vector of clusters
BC	CV^E	CV of event distribution		
DBC	$\bar{t}$	Avg. interaction time		
DBC	σ_t	Std. Dev. of interaction times		
DBC	$\log(T)$	Test statistic for $\bar{t}$		
DP	JSD	Differences in daily behaviour		

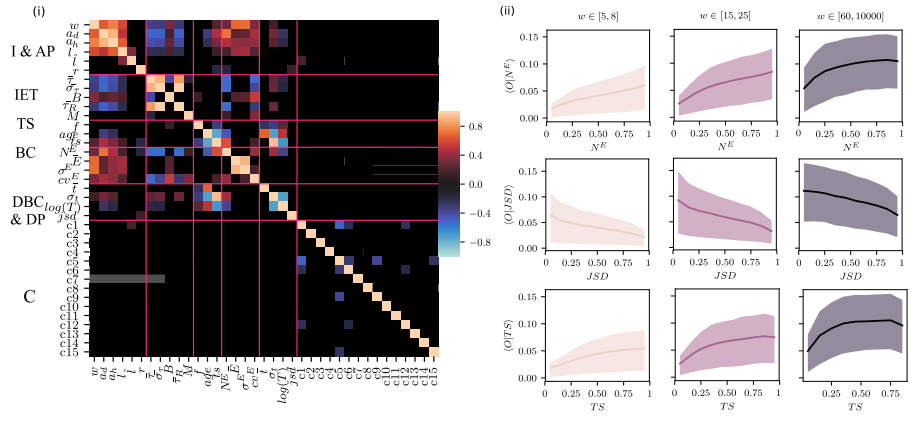

Figure 4 Exploratory analysis of features of human communication: (*i*) Feature correlation matrices measured by Pearson's correlation coefficient. Features are sorted according to their modelling approach, each case divided by pink lines. Features display high within-group correlations, with lower between-group correlations. Weekly clusters (last group) show no relevant correlations among themselves or to other variables, with two main exceptions: negative correlations between clusters $C1$ (late night) and $C5$ (weekday worktimes), and clusters $C5$ and $C9$ (weekend night). (*ii*) Average topological overlap given the ranks of three variables correcting for three different levels of communication intensity (w), with the shaded area depicting 80% of the distribution. From top to bottom: number of bursty trains (N^E), Jensen-Shannon Divergence for difference in daily patterns (JSD), temporal stability (TS). Variable rankings are normalized to be on the [0, 1] interval

three features—the number of bursty trains (N^E), the daily pattern difference (JSD), and the temporal stability TS—in Fig. 4 (ii). It is clear that these features correlate with overlap even within groups of ties with a narrow intensity range; this holds for other measures of communication intensity and other features (IET, etc) as well. See SI for further details.

For our predictive task, we applied machine-learning models (see below) to two different scenarios: a) using each feature as a single predictor, b) using each feature along with the best-performing features in the previous task, we include an additional scenario using the full set of features in the SI. These scenarios allow us i) to identify the individual features that encode most information on overlap and ii) to compare the performance of these features with commonly used measures and see how complementary they are.

As there is no natural scale for overlap that would relate it to the latent tie strengths, we take a nonparametric approach and focus on predicting overlap rank instead of overlap values. The prediction problem itself was transformed into the binary decision problem of predicting high/low overlap values. We selected a range of high/low overlap values according to the overall distribution in our population, with cutoff points every fifth percentile, $\eta(\hat{O}^t)$. As such, the goal of each binary prediction problem is to map the values where each feature might gain importance, and thus most of most of the prediction problems have unbalanced training data over the overlap range. This however, allows us to pose predictive problems without additional assumptions on the response variable. For completeness, we include results with balanced training data via down-sampling in the SI, where we find a similar results in terms of variable rankings, albeit with better extreme-case performance. For each scenario, we ran four machine-learning models, which we average in order to avoid any model-specfic shortcomings: logistic regression (LR), random forests (RF), quadratic discriminant analysis (QDA) and AdaBoost classifier (ABC). For the CDR data, we obtained a sample of 500,000 ties, performed 3-fold cross-validation for our overlap prediction tasks, and measured the predictive performance of our models via Matthews Correlation Coefficient (MCC) [64], a classification performance metric for binary data related to Person's correlation coefficient, and used for it's ability to handle imbalanced and asymmetric data [65].

The predictive performance of all individual features is shown in Fig. 5. Results are shown for the averaged overlap, $\hat{O}^t$. For static overlap O, see SI. In addition, we include $C^* = (C_1, \ldots, C_{15})$, the vector of cluster weights for a tie's weekly call profile. Although C^* is not a single variable, we include it as a means of comparing how much information is encoded by the weekly call profile.

On average, nine features outperform the number of calls w in predicting topological overlap: the number of days a_d and the number of hours a_h with calls, the number of bursty trains N^E, temporal stability TS, the weekly call profile C^*, three features of the distribution of bursty cascades (DBC), and tie *age*. When comparing with the baseline, however, only the first four features consistently show a higher predictive capacity than w. In other cases, a feature's performance varies might be higher or lower to w's depending on the overlap range.

The performance of predictors differs for low or high overlap cutoff percentiles $\eta(\hat{O}^t)$, which is indicative of how these measures perform overall: a_d, a_h and N^E encompass a broad spectrum of values centered around the median of the overlap distribution. The weekly call profile C^* has a wider spectrum and is one of the few features with nonzero MCC for all percentiles, even though its predictive performance for mid-range percentiles is smaller than that of the three top-ranking features. The features TS and DBC (TS, *age*, σ_t, $\log(T)$ and $\bar{t}$) tend to have higher predictive performance skewed towards smaller $\eta(\hat{O}^t)$ values. Note, that these results depict deliberately unbalanced datasets for small and low

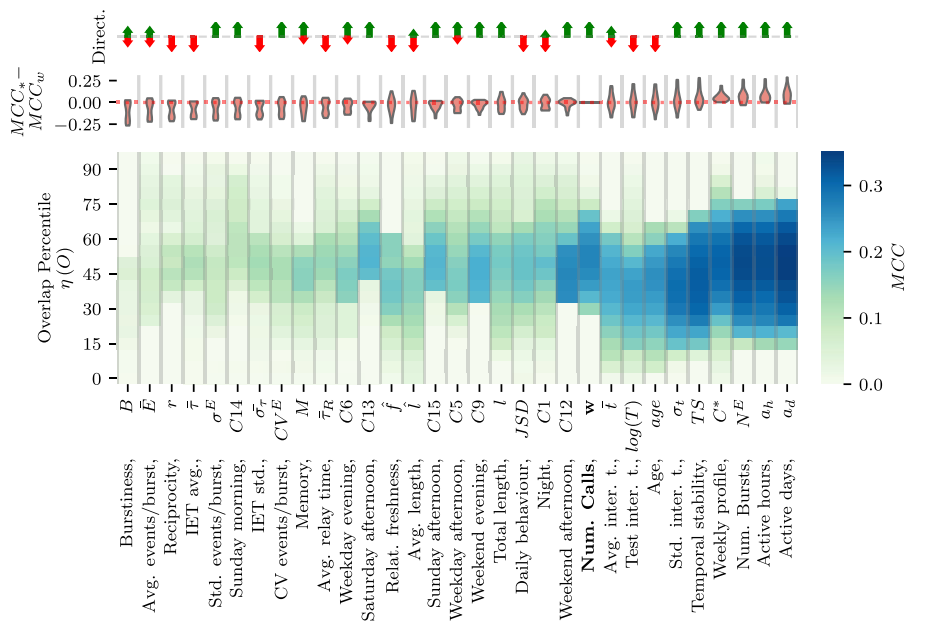

Figure 5 Results of binary overlap prediction using single features, where the x-axis represents features, and y-axes depict: (*top*) Directionality of feature association, where green upwards arrows represent a positive association between the feature and the averaged overlap, red downwards arrows represent a negative association, and both green and red arrows represents non-linear association. (*middle*) Feature comparison with the baseline communication intensity MCC ($\mathbf{MCC}_* - \mathbf{MCC}_w$, where $*$ is a feature) across all overlap cutoff values. A distribution fully above the red line implies a feature completely outperforms w. (*bottom*) Maximum MCC for four models trained with single-feature predictors, where each variable is used to predict static overlap using RF, ABC, LG and QDA. The color represents the maximum **MCC** over the four ML models. Variables are ranked according to their average performance over all cutoff values. For visualization purposes, we exclude weekly profile variables with an average performance of less than MCC = 0.05

cutoff values. Using down-sampling techniques when training models allows the feature's predictive range to extend to a broader spectrum of overlap values.

The fraction of contacts in some component clusters of the weekly call profile is surprisingly predictive of overlap. In particular the weekend day cluster alone (C_{12}) has a high predictive performance for mid-range values of $\eta(\hat{O}^t)$. The cluster for early morning and weekday nights (C_1) also ranks highly for average overlap prediction. In this case, correlation with overlap was mostly negative, suggesting that a high fraction of calls at certain times might indicate weak ties. We provide a more complete analysis of the predictive power and the importance of the different components of C^* in the SI.

We include the directionality of the association where, for example, a larger number of active days (a_d) implies a larger overlap. Some features have non-linear associations, such as the average interaction time over the observation window ($\bar{t}$); in this case, the average overlap peaks at central average interaction times, and decreases towards the beginning and end of the observation window.

On Fig. 6 we compare the effect of including additional information on the prediction task by using pairs of variables as predictors (F, X), where F is one of the three best-performing features (a_d, a_h, N^E) or the number of calls (w), and X is the set of all other features. These variables' performance increases moderately when used in tandem when compared with the baseline single predictor, with an average increase of 16.8% for (a_d, X) against a_d. Many features are highly correlated, which explains the small performance

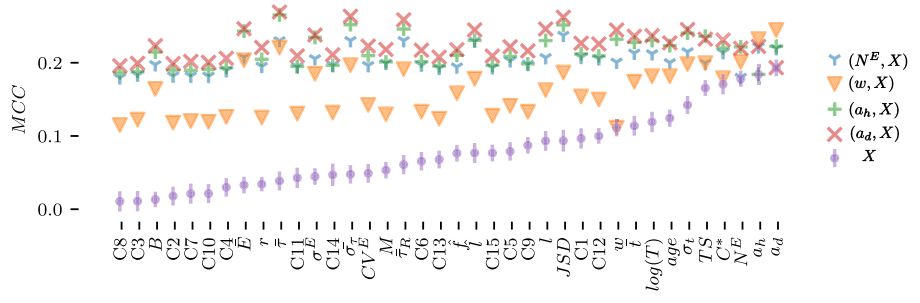

Figure 6 Comparison between single and dual-variable models (F, X) for $F = N^E, w, a_h, w_d$, where we depict the average MCC performance over all overlap cutoff values. For the single-variable case (X), error bars are 2 standard deviations, obtained via bootstrapping

boost. For a small set of features, however, the average performance increases considerably, up to 39.5% for $(a_d, \bar{\sigma})$. Notably, the compound effect of feature pairs is higher with variables that have low single-feature predictive performance. This includes variables derived from the IET, such as $\bar{\tau}$, $\bar{\sigma}_\tau$, differences in daily patterns JSD, and features of call duration, l and $\hat{l}$. In the SI, we include an additional analysis for predicting overlap using the full set of features, which we consider to be an upper bound on the MCC for these variables. In this case, the maximum MCC equals to 0.45.

4.3 Analysis of additional datasets

We performed the single-variable analysis for our additional datasets. To do so, we adapted our measures according to the main communication channel, so that the number of contacts (w) is now in terms of the number of emails (Enron), text messages (Copenhagen) and wall or thread postings (FB and Forum, respectively); we discarded features that did not have a direct translation such as total call time. We do not include the weekly activity profiles since most datasets do not contain sufficient data, missing either specific date information or enough ties to use our clusterization process. Also, we use static overlap O as a predictive variable. As before, we performed 3-fold cross-validation and evaluate results with MCC.

Figure 7 depicts the results of our four additional datasets. We find the number of active periods to still rank highly in the Forum and FB datasets, while displaying a mid-range performance for Enron and Copenhagen. Noticeably, features of temporal stability and the distribution of bursty cascades (TS, age and σ_t) remain ranking highly for predicting overlap. A relevant difference in performance comes from JSD, the differences in daily behaviour, which ranks highly for the four datasets. This could be an effect of the network capturing specific foci (see Discussion) In addition, there is a stark difference in the performance across quantiles for the Enron and Copenhagen networks. Particularly for the latter, MCC does displays larger variability between adjacent percentiles. This is likely to be an effect the small sample size in this network, where the small number of edges results in an uneven overlap distribution.

5 Discussion

Human communication patterns encode information on their local network topology. In this paper, we conceptualized tie strength as a latent variable that manifests independently

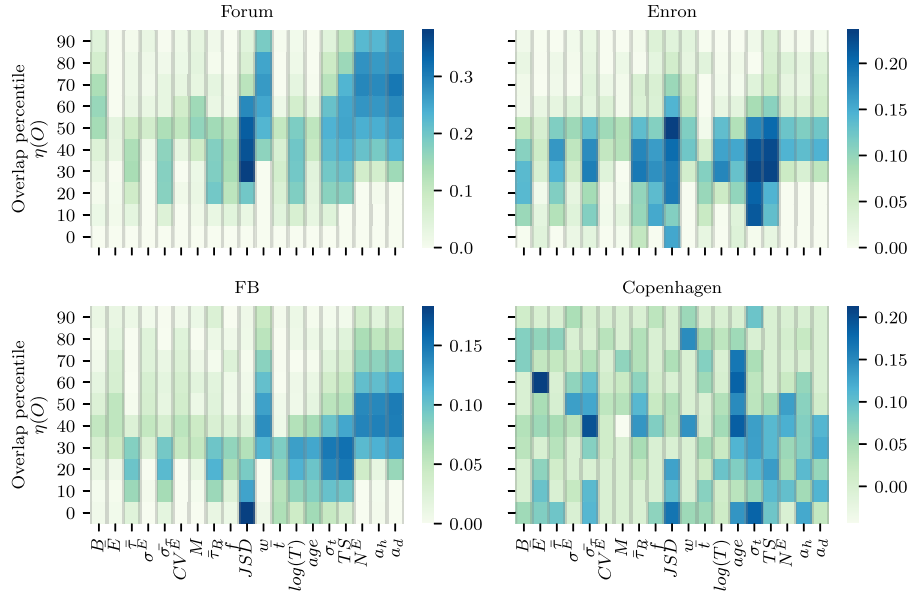

Figure 7 Results for four additional datasets, where each table depicts the MCC values for temporal features of binary classification over the overlap distribution. Features are sorted in the same order as Fig. 5 (*bottom*). Datasets: (*top left*) Forum, (*top right*) Enron, (*bottom left*) Facebook, (*bottom right*) Copenhagen

as both local network topology and as patterns of communication between two nodes. We identified which features of these dyadic interactions are the best predictors for the neighbourhood overlap of a tie and therefore for the latent tie strength. We find that while commonly-used aggregated measures such as the total number of calls are adequate indicants of network overlap, our results show that alternative proxy measures contain information not captured by mere intensity features. We focused on quantifying different temporal aspects of human communication, using both sequential and cyclical features, and we assumed this topological tie strength to be constant during the observation period; for linking dynamic topology with temporal features, see, e.g., [31]. We showed that several of these distinct approaches capture information on network topology, results which also stand across different communication foci.

The number of days and hours with contacts (a_d and a_h, respectively) outperformed all other variables in our main prediction task, as did the number of bursty cascades N^E. Notably, these variables are conceptually similar to features measuring communication intensity, but with the key difference that part of bursty behaviour is removed through temporal aggregation. In addition to these, simple variables related to the time of the first and last communication (TS and *age*) performed better than the communication-intensity features. These variables ranked highly both in the main dataset and in the smaller datasets.

We introduced a weekly call profile C^* and found it to be highly informative on the neighbourhood overlap of ties. Notably, even though C^* was not the best predictor, it had the highest predictive power for the widest range of overlap cutoff values, providing a richer characterization that other features. Interestingly, C^*'s performance does not increase significantly in combination with new variables, which might suggest that the weekly profile contains information on intensity as well. A simple mechanism for encoding a large number of communication events could be through several active clusters, for

example. What is more, we found strong evidence that individual calling times during the week convey information on network topology. Notably, for our dataset in a European country, weekend afternoons proved to have a higher correlation with overlap than most other variables, whereas weekday nights and early mornings were associated with low overlap. These results pave the way for interesting lines of research. For example, one can use different data sets to compare the differences of weekdays and times of days across contexts and cultures.

In the case of modelling bursty trains, the parameter Δt determines the period where two calls are considered to be correlated. Previous research had found that the distribution of calls within trains did not vary significantly with different Δt values [47]. Although we did find differences in predictive performance, which included an optimal value of $\Delta t^* = 26$ hours, we also found evidence that a wide range of Δt values outperformed w. This suggests that in practical applications, the aggregation of temporally correlated calls might already improve the topological information encoded in the variable.

Measuring differences in daily call patterns (JSD) also proved to encompass topological information, an effect more evident when predicting static overlap (see SI) and dynamic overlap in our dual-variable scenario. For the additional datasets, JSD was either the best or the one of the best-performing features. This was slightly surprising, as the relationship to network topology is not as straightforward as other features. We hypothesized two possible explanations for this, which are not mutually exclusive. In the first case, there could be a latent homophilic effect, where activity encodes information on, *e.g.*, age or work relations. A second possible explanation is that strong ties engage in correlated events, where person A's contact is followed by the person B's contact. Despite the strong association, further research is needed to uncover the drivers of this relationship. The use of temporal stability also provides a useful characterization, as it is one of the most simple features that only requires two observations. Indeed, we do not delve into the effect of the observation window into the use of this variable, where tie decay is more likely to occur, along with the topological changes it implies [31, 39].

We found that these features were relevant in our main CDR dataset, but also in other communication channels and in specific social networks. As in the main dataset, we found that counts of days and hours, measures of temporal stability, along with the differences in daily patterns, were the best predictors for the strength of ties. Granted, a relevant criteria for these results was that the Granovetter effect was present for the contact counts. Whether these results hold for different cases would require a more careful analysis of each social network in question, and whether the underlying Granovetter hypothesis about strong ties and overlapping circles of friends would be valid.

If one needs to pick a single simple measure for tie strength based on this study it would be the number of days with contact. However, this measure would have only about two thirds of the predictive power as compared to using the full contact sequence (when measured with MCC to predict $\hat{O}^t$). That is, the latent tie strength is a combination of multiple features which reflects the different facets of human relationships. Our results suggest that such important facets include regularity of contact, total amount of time spent, and the type of interaction reflected by the time and weekday of the contact.

We should also note here that we did not investigate the direction of causality, but only the association of variables. That is, we do not answer the question of if high overlap values are followed by high latent tie strengths or the other way around. If each feature represents

different aspects of the latent tie strength then one could also study each of them separately as predictors of overlap in the future or vice versa. Moreover, our results might be dependent on cultural features, communication medium, technology and other variables, and thus might not be directly transferable to other data sets. However, if one has access to a social network based on contact events, then it is straightforward to use the framework we have set up here and find the features which are most important in a specific context.

Lastly, the list of features we constructed here is by no means exhaustive and it is based on the current literature on analysing temporal social networks. However, our framework provides a way to benchmark any new features as an independent predictor of the latent tie strength, or as an additional facet of the tie strength by inspecting its performance together with other features.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-020-00256-5.

Additional file 1. Supplementary information is a file in PDF format that contains eight sections detailing further information and tests that validate our results. It includes an extended analysis of some of our features, repeating the main experiment using static overlap and balanced training data, a breakdown of our results by ML model, additional performance metrics, analyzing the effect of Δt on bursty cascades and using weekly signatures for predicting overlap. (PDF 2.2 MB)

Acknowledgements
The authors acknowledge the computational resources provided by the Aalto Science-IT project.

Funding
JS acknowledges funding from the Academy of Finland, project n:o 297195. MK acknowledges funding from the Academy of Finland, project n:o 320781.

Abbreviations
ABC, AdaBoost Classifier; AP, Active Periods; BC, Bursty Cascades; CDRs, Call Detail Records; DBC, Distribution of Bursty Cascades; DP, Daily Patters; I, Intensity; IET, Inter-Event Time; JSD, Jensen-Shannon Divergence; LR, Logistic Regression; MCC, Matthews Correlation Coefficient; ML, Machine Learning; QDA, Quadratic Discriminant Analysis; RF, Random Forests; SI, Additional file 1; TS, Temporal Stability.

Availability of data and materials
The CDR dataset analysed during the current study is not publicly available due to a signed non-disclosure agreement. The dataset contains sensitive information of the subscribers. The four remaining datasets are openly available at their marked reference.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Designed the study: JUC, MK. Analyzed the data: JUC. Wrote the paper: JUC, MK, JS. All authors read and approved the final manuscript.

Publisher's Note

Received: 6 July 2020 Accepted: 3 December 2020 Published online: 14 December 2020

References

1. Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási A-L (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci 104(18):7332–7336. https://doi.org/10.1073/pnas.0610245104
2. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási A-L (2008) Uncovering individual and collective human dynamics from mobile phone records. J Phys A, Math Theor 41(22):224015. https://doi.org/10.1088/1751-8113/41/22/224015
3. Saramäki J, Moro E (2015) From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. Eur Phys J B 88(6). https://doi.org/10.1140/epjb/e2015-60106-6

4. Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895. https://doi.org/10.1126/science.1165821
5. Karsai M, Kivelä M, Pan RK, Kaski K, Kertész J, Barabási A-L, Saramäki J (2011) Small but slow world: how network topology and burstiness slow down spreading. Phys Rev E 83(2). https://doi.org/10.1103/physreve.83.025102
6. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958
7. Li T, Dejby J, Albert M, Bengtsson L, Lefebvre V (2019) Estimating the resilience to natural disasters by using call detail records to analyse the mobility of internally displaced persons. arXiv:1908.02381. https://doi.org/10.5281/zenodo.3349848
8. Altuncu MT, Kaptaner AS, Sevencan N (2019) Optimizing the access to healthcare services in dense refugee hosting urban areas: a case for Istanbul. arXiv:1903.09614
9. Zeller RA, Nock SL, Carmines EG (1982) Measurement in the social sciences: the link between theory and data. Contemp Sociol 11(1):79. https://doi.org/10.2307/2066656
10. Marsden PV, Campbell KE (1984) Measuring tie strength. Soc Forces 63(2):482. https://doi.org/10.2307/2579058
11. Marsden PV, Campbell KE (2012) Reflections on conceptualizing and measuring tie strength. Soc Forces 91(1):17–23. https://doi.org/10.1093/sf/sos112
12. Karsai M, Perra N, Vespignani A (2014) Time varying networks and the weakness of strong ties. Sci Rep 4(1). https://doi.org/10.1038/srep04001
13. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. EPJ Data Sci 1(1). https://doi.org/10.1140/epjds4
14. Kovanen L, Kaski K, Kertesz J, Saramaki J (2013) Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. Proc Natl Acad Sci 110(45):18070–18075. https://doi.org/10.1073/pnas.1307941110
15. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4(1). https://doi.org/10.1140/epjds/s13688-015-0046-0
16. Park PS, Blumenstock JE, Macy MW (2018) The strength of long-range ties in population-scale social networks. Science 362(6421):1410–1413. https://doi.org/10.1126/science.aau9735
17. Miritello G, Lara R, Moro E (2013) Time allocation in social networks: correlation between social structure and human communication dynamics. Springer, Berlin, pp 175–190. https://doi.org/10.1007/978-3-642-36461-7_9.
18. Saramaki J, Leicht EA, Lopez E, Roberts SGB, Reed-Tsochas F, Dunbar RIM (2014) Persistence of social signatures in human communication. Proc Natl Acad Sci 111(3):942–947. https://doi.org/10.1073/pnas.1308540110
19. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380. https://doi.org/10.1086/225469
20. Wuchty S (2009) What is a social tie? Proc Natl Acad Sci 106(36):15099–15100. https://doi.org/10.1073/pnas.0907905106
21. Wellman B (1990) Studying personal communities. Soc Netw 26972. https://doi.org/10.1371/journal.pone.0026972
22. Granovetter M (1985) Economic action and social structure: the problem of embeddedness. Am J Sociol 91(3):481–510. https://doi.org/10.1086/228311
23. Friedkin NE (1990) A guttman scale for the strength of an interpersonal tie. Soc Netw 12(3):239–252. https://doi.org/10.1016/0378-8733(90)90007-v
24. Brewer DD (2000) Forgetting in the recall-based elicitation of personal and social networks. Soc Netw 22(1):29–43. https://doi.org/10.1016/s0378-8733(99)00017-9
25. Wilmot W (1987) Dyadic communication. ISBN 0394358260
26. Dindia K, Canary DJ (1993) Definitions and theoretical perspectives on maintaining relationships. J Soc Pers Relatsh 10(2):163–173. https://doi.org/10.1177/026540759301000201
27. Hallinan MT (1978) The process of friendship formation. Soc Netw 1(2):193–210. https://doi.org/10.1016/0378-8733(78)90019-9
28. Burt RS (2000) Decay functions. Soc Netw 22(1):1–28. https://doi.org/10.1016/s0378-8733(99)00015-5
29. Burt RS (2002) Bridge decay. Soc Netw 24(4):333–363. https://doi.org/10.1016/s0378-8733(02)00017-5
30. Backstrom L, Kleinberg J (2014) Romantic partnerships and the dispersion of social ties. In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing. CSCW, vol 14. https://doi.org/10.1145/2531602.2531642
31. Navarro H, Miritello G, Canales A, Moro E (2017) Temporal patterns behind the strength of persistent ties. EPJ Data Sci 6(1). https://doi.org/10.1140/epjds/s13688-017-0127-3
32. Ayres J (1983) Strategies to maintain relationships: their identification and perceived usage. Commun Q 31(1):62–67. https://doi.org/10.1080/01463378309369487
33. Feld SL (1981) The focused organization of social ties. Am J Sociol 86(5):1015–1035. https://doi.org/10.1086/227352
34. Kahanda I, Neville J (2009) Using transactional information to predict link strength in online social networks. In: Third international AAAI conference on weblogs and social media.
35. Miritello G (2013) Temporal patterns of communication in social networks. https://doi.org/10.1007/978-3-319-00110-4.
36. Holme P, Saramäki J (2012) Temporal networks. Phys Rep 519(3):97–125. https://doi.org/10.1016/j.physrep.2012.03.001
37. Wang D, Pedreschi D, Song C, Giannotti F, Barabasi A-L (2011) Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining - KDD. https://doi.org/10.1145/2020408.2020581
38. Raeder T, Lizardo O, Hachen D, Chawla NV (2011) Predictors of short-term decay of cell phone contacts in a large scale communication network. Soc Netw 33(4):245–257. https://doi.org/10.1016/j.socnet.2011.07.002
39. Miritello G, Moro E, Lara R, Martínez-López R, Belchamber J, Roberts SGB, Dunbar RIM (2013) Time as a limited resource: communication strategy in mobile phone networks. Soc Netw 35(1):89–95. https://doi.org/10.1016/j.socnet.2013.01.003
40. Wuchty S, Uzzi B (2011) Human communication dynamics in digital footsteps: a study of the agreement between self-reported ties and email networks. PLoS ONE 6(11):26972. https://doi.org/10.1371/journal.pone.0026972
41. Gilbert E, Karahalios K (2009) Predicting tie strength with social media. In: Proceedings of the 27th international conference on human factors in computing systems - CHI 09. https://doi.org/10.1145/1518701.1518736

42. Karimi F, Ramenzoni VC, Holme P (2014) Structural differences between open and direct communication in an online community. Phys A, Stat Mech Appl 414:263–273. https://doi.org/10.1016/j.physa.2014.07.037
43. Rossi RA, Ahmed NK (2015) The network data repository with interactive graph analytics and visualization. In: AAAI. http://networkrepository.com
44. Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in Facebook. In: Proceedings of the 2nd ACM workshop on online social networks - WOSN 09. ACM, New York. https://doi.org/10.1145/1592665.1592675.
45. Sapiezynski P, Stopczynski A, Lassen DD, Lehmann S (2019) Interaction data from the copenhagen networks study. Sci Data 6(1). https://doi.org/10.1038/s41597-019-0325-x
46. Heydari S, Roberts SG, Dunbar RIM, Saramäki J (2018) Multichannel social signatures and persistent features of ego networks. Appl Netw Sci 3(1). https://doi.org/10.1007/s41109-018-0065-4
47. Karsai M, Kaski K, Barabási A-L, Kertész J (2012) Universal features of correlated bursty behaviour. Sci Rep 2(1). https://doi.org/10.1038/srep00397
48. Goh K-I, Barabási A-L (2008) Burstiness and memory in complex systems. Europhys Lett 81(4):48002. https://doi.org/10.1209/0295-5075/81/48002
49. Barabási A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435(7039):207–211. https://doi.org/10.1038/nature03459
50. Jo H-H, Hiraoka T, Kivelä M (2019) Burst-tree decomposition of time series reveals the structure of temporal correlations. arXiv:1907.13556 [physics.data-an]
51. Kivelä M, Porter MA (2015) Estimating interevent time distributions from finite observation periods in communication networks. Phys Rev E 92(5). https://doi.org/10.1103/physreve.92.052813
52. Kivelä M, Pan RK, Kaski K, Kertész J, Saramäki J, Karsai M (2012) Multiscale analysis of spreading in a large communication network. J Stat Mech Theory Exp 2012(03):03005. https://doi.org/10.1088/1742-5468/2012/03/p03005
53. Holme P, Liljeros F (2015) Birth and death of links control disease spreading in empirical contact networks. Sci Rep 4:4999. https://doi.org/10.1038/srep04999
54. Panda S, Hogenesch JB, Kay SA (2002) Circadian rhythms from flies to human. Nature 417(6886). https://doi.org/10.1038/417329a
55. Aledavood T, Lehmann S, Saramäki J (2018) Social network differences of chronotypes identified from mobile phone data. EPJ Data Sci 7(1). https://doi.org/10.1140/epjds/s13688-018-0174-4
56. Wittmann M, Dinich J, Merrow M, Roenneberg T (2006) Social jetlag: misalignment of biological and social time. Chronobiol Int 23(1–2):497–509. https://doi.org/10.1080/07420520500545979
57. Aledavood T, Lehmann S, Saramäki J (2015) Digital daily cycles of individuals. Front Phys 3:73. https://doi.org/10.3389/fphy.2015.00073
58. Aledavood T, López E, Roberts SGB, Reed-Tsochas F, Moro E, Dunbar RIM, Saramäki J (2015) Daily rhythms in mobile telephone communication. PLoS ONE 10(9):0138098. https://doi.org/10.1371/journal.pone.0138098
59. Aledavood T, López E, Roberts SGB, Reed-Tsochas F, Moro E, Dunbar RIM, Saramäki J (2015) Channel-specific daily patterns in mobile phone communication. arXiv:1507.04596
60. Adan A, Archer SN, Hidalgo MP, Milia LD, Natale V, Randler C (2012) Circadian typology: a comprehensive review. Chronobiol Int 29(9). https://doi.org/10.3109/07420528.2012.719971
61. Aledavood T, Kivimäki I, Lehmann S, Saramäki J (2020) A non-negative matrix factorization based method for quantifying rhythms of activity and sleep and chronotypes using mobile phone data. arXiv:2009.09914
62. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27(1):415–444.
63. Asikainen A, Iñiguez G, Ureña-Carrión J, Kaski K, Kivelä M (2020) Cumulative effects of triadic closure and homophily in social networks. Sci Adv 6(19):7310.
64. Matthews BW (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochim Biophys Acta, Protein Struct 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9
65. Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using matthews correlation coefficient metric. PLoS ONE 12(6):0177678. https://doi.org/10.1371/journal.pone.0177678