Supplementary Information for

Going beyond communication intensity features for estimating tie strengths in social networks

J. Ureña-Carrión*, J. Saramäki, and M. Kivelä

*Corresponding author email: javier.urenacarrion@aalto.fi

Contents

S1 Analysis of Variables

S1.1 Static and Dynamic Overlap

In this paper, we find proxies for measuring the strength of ties in communication networks by claiming that tie strength manifests both in communication patterns and network topology. We measure embeddedness via average overlap over 1-month aggregation windows. More explicitly, the overlap obtained over a shorter time period ΔT and using a sliding window approach as to obtain a time series of overlap values $\{O_i\}$. In our paper, we use a $\Delta T = 1$ month as an aggregation period, and here we compare this average dynamic measure to overlap measured on the full aggregation window.

Figure 1 displays the joint distribution of static and dynamic overlap measures. Both network statistics have similar distributions, with Pearson's correlation coefficient of 0.77 and a rank correlation of 0.81. The main difference between is the larger number of zero-valued entries for dynamic overlap, up to 5.18% as opposed to static overlap's 1.03%. We briefly see how the set of common neighbors ($\mathcal{N}_i \cup \mathcal{N}_j$) changes from static to dynamic overlap. We divide the set of common neighbors onto three cases depending on their stability across the dynamic overlap's aggregation period. We find that only 1.32% of neighbors were connected to both nodes in a tie at all aggregation windows, 50.66% neighbors were connected to both nodes during at least one aggregation window and 48.02% of neighbors not connected to both nodes during the same aggregation period. This high turnover of common neighbors helps explain why dynamic overlap tends to have smaller values than it's static counterpart, since at least 48% of neighbors from O do not translate to common neighbors on $\hat{O}^t$.

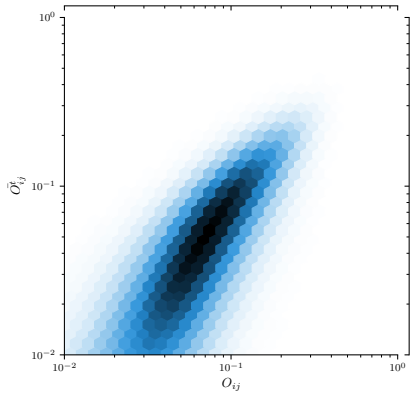

Figure S1. Joint distribution of static O and dynamic $\hat{O}^t$ overlap. For visualization purposes we only include positive overlap values.

S1.2 Creation of weekly cluster profiles

Our analysis of weekly activity profiles involves a two-step procedure where we first have a high-granularity approach by dividing the week into $n = 168$ hours, followed by a clustering process to reduce the number of variables. There is a trade-off between the number of variables and the information they contain on overlap—a large number of variables implies that for most ties, their weekly profiles will be

zero-valued, whereas having a low number of variables (e.g., two weekly profile variables of calls placed on work and leisure times) might ignore valuable details on which topological information is encoded at certain times.

To determine our clusters, we selected a sample of 100,000 ties and for each one obtained 168 weekly activity variables $\{\phi^h\}_{h=1...168}$, where ϕ^h contains the fraction of calls placed at hour h. We then computed the correlation matrix to detect timing where behaviour is similar. We used Markov Cluster Algorithm (MCA) [1], a method that uses the correlation matrix as input, as well as a parameter ψ that determines a cutoff value that determines the granularity of the clusters (which we denote by C_ψ). Given ψ, we can determine the weekly clusterized profiles $\{\phi^c\}_{c\in C_\psi}$, and we then determine the smallest cutoff value that captures as much diversity in overlap values as our high-granularity approach, our criteria for clusterization.

S1.3 Relationship to overlap correcting for w

We present plots Figures 2, 3, 4 that depict relationships of the type $\langle O|F\rangle$, for all features F. Since one of the main goals of this research is to understand whether the features provide additional information not captured by communication intensity w; we correct for part of the effect of w by plotting $\langle O|F\rangle$ at five different levels of w, determined by equal-sized quantiles.

We observe that for a large set of features an increase in F implies changes in both the average overlap and the overlap distribution. These figures provide evidence that for most features the relationship to overlap is not necessarily linear. Indeed, our features seem to both interact with communication intensity w and encode different information about overlaps at different values. For instance, for the number of days and hours with contacts (a_d and a_h on Figure 2), the average overlap increases almost linearly with $F = a_d, a_h$ for communication intensity $w \leq 25$; overlap increases at a decreasing rate with $F = a_d, a_h$ for higher communication intensity. Now, considering the average IET ($\bar{\tau}$ on Figure 2), we find overlap to be more sensitive to changes in $\bar{\tau}$ when it is smaller, and after a certain degree $\bar{\tau}$ becomes less effective at encoding information of overlap. Most relationships seems to be both non-trivial and non-linear, we expect ML models that favour both variable interactions and non-linearities to be more efficient at capturing overlap.

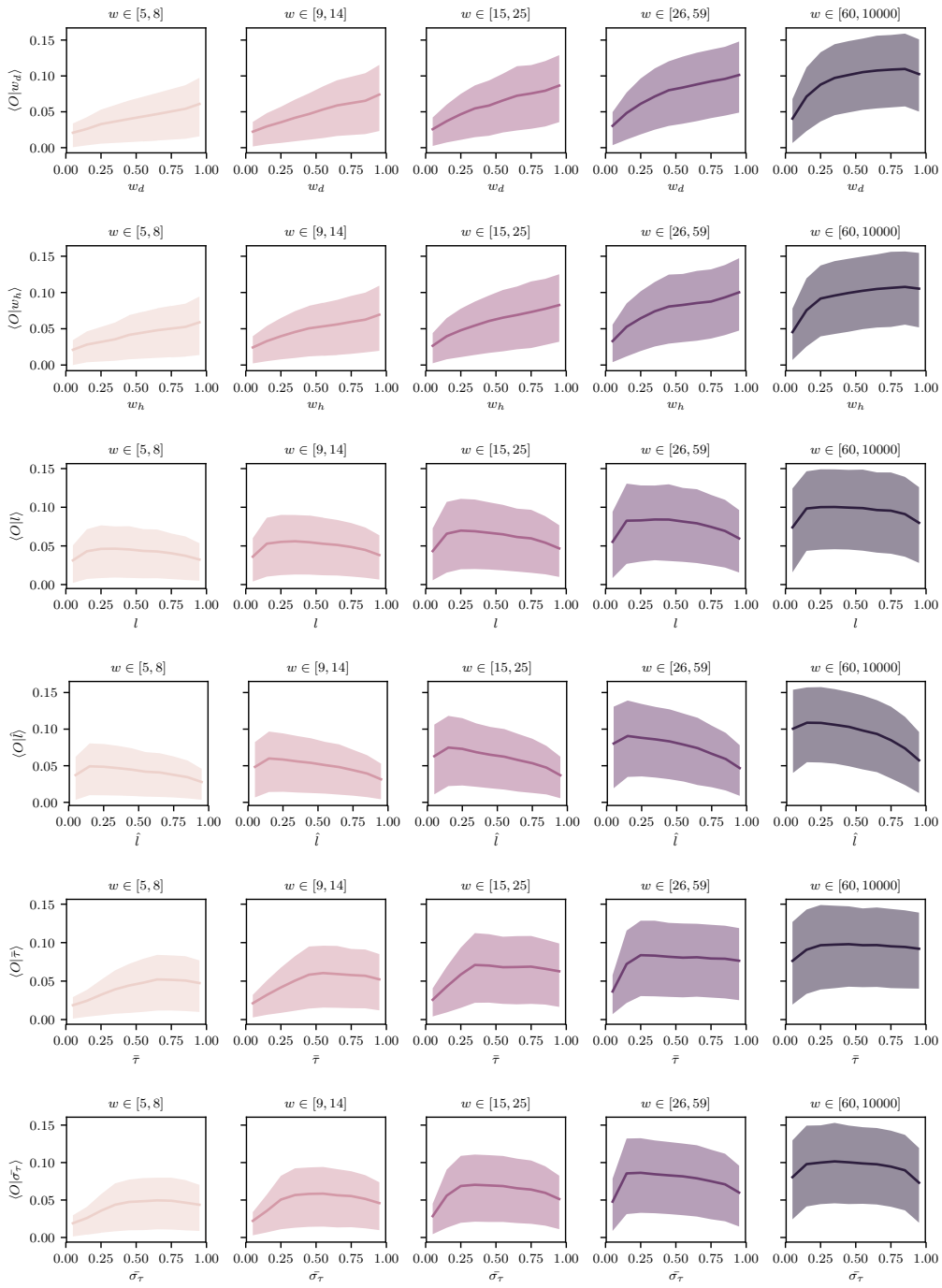

Figure S2. Average topological overlap for non-cluster variables correcting for different levels of communication intensity (w, in columns). Shade includes 80% of the distribution. From top to bottom: number of days with contacts (a_d), number of hours with contacts (a_h), total call length or duration (l), average call length ($\hat{l}$), average IET ($\bar{\tau}$), standard deviation of the IET ($\bar{\sigma}_\tau$).

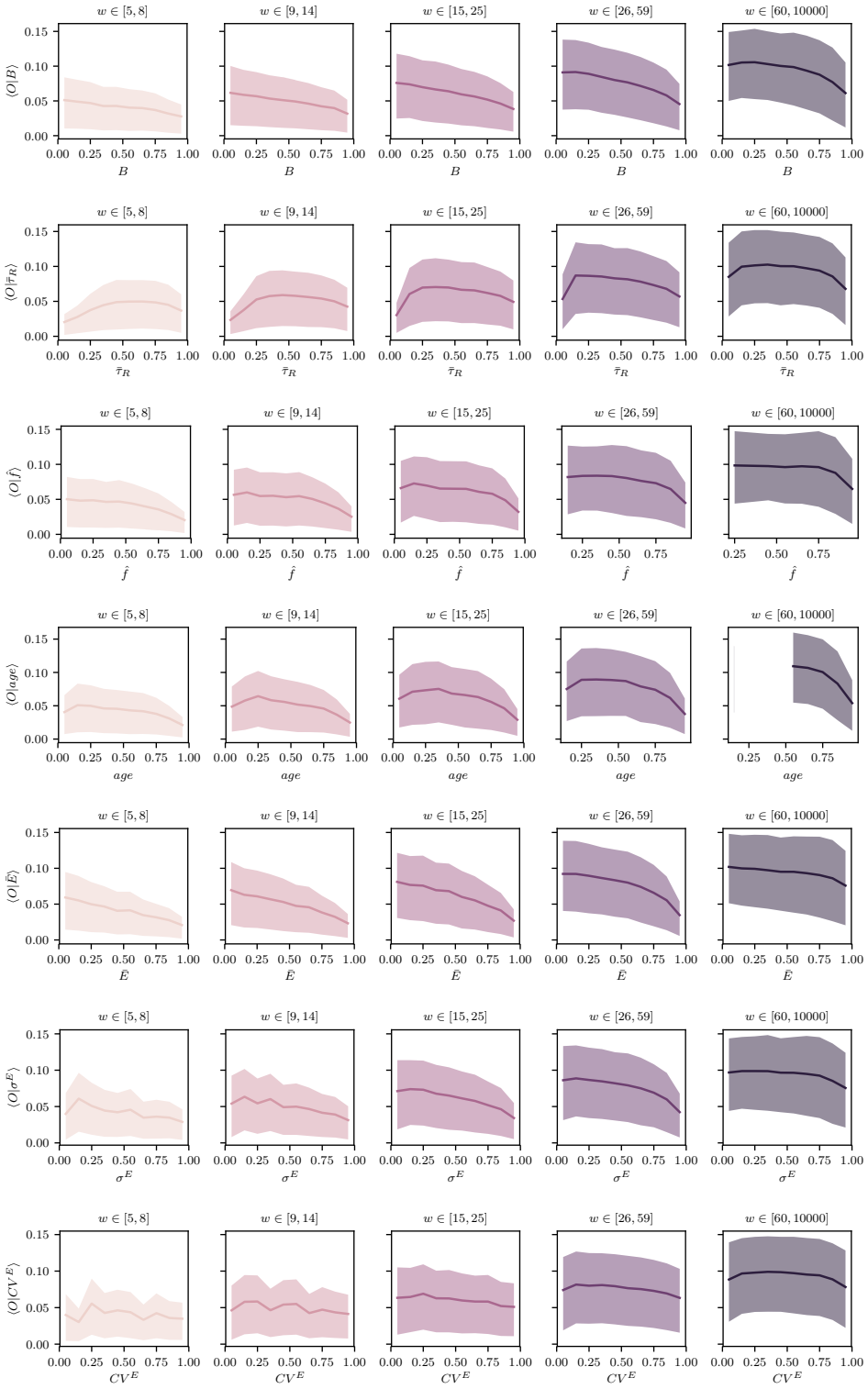

Figure S3. Average topological overlap for non-cluster variables correcting for different levels of communication intensity (w, in columns). Shade includes 80% of the distribution. From top to bottom: burstiness (B), average inter-relay time ($\bar{\tau}_R$), relative freshness ($\hat{f}$), tie age (age), average number of calls per bursty train ($\bar{E}$), std. deviation of number of calls per bursty train (σ^E), coefficient of variation of E (CV^E).

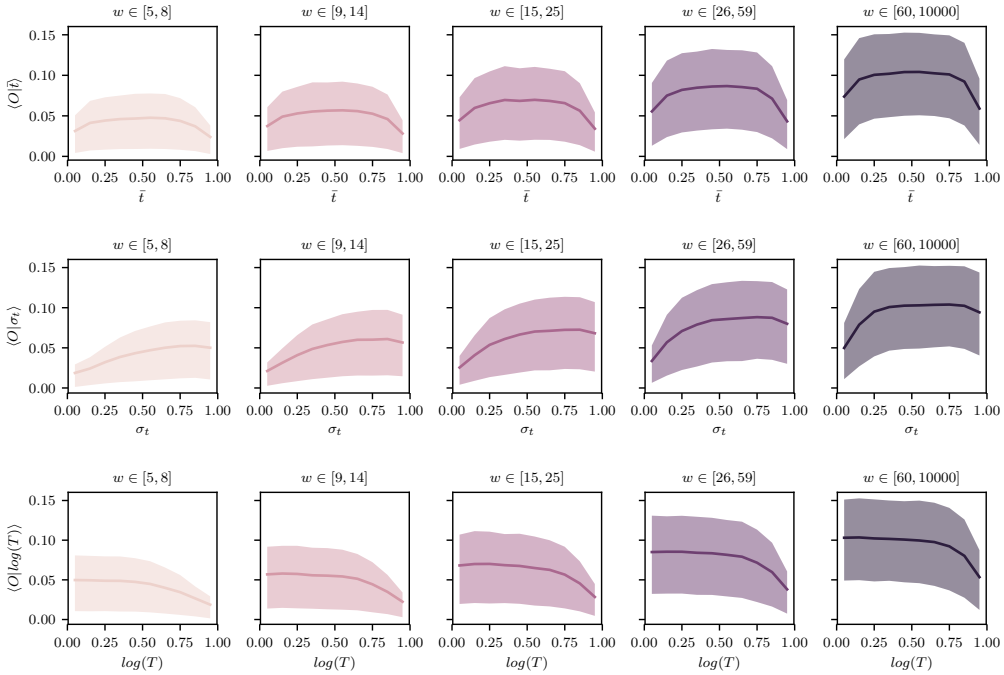

Figure S4. Average topological overlap for non-cluster variables correcting for different levels of communication intensity (w). From top to bottom:

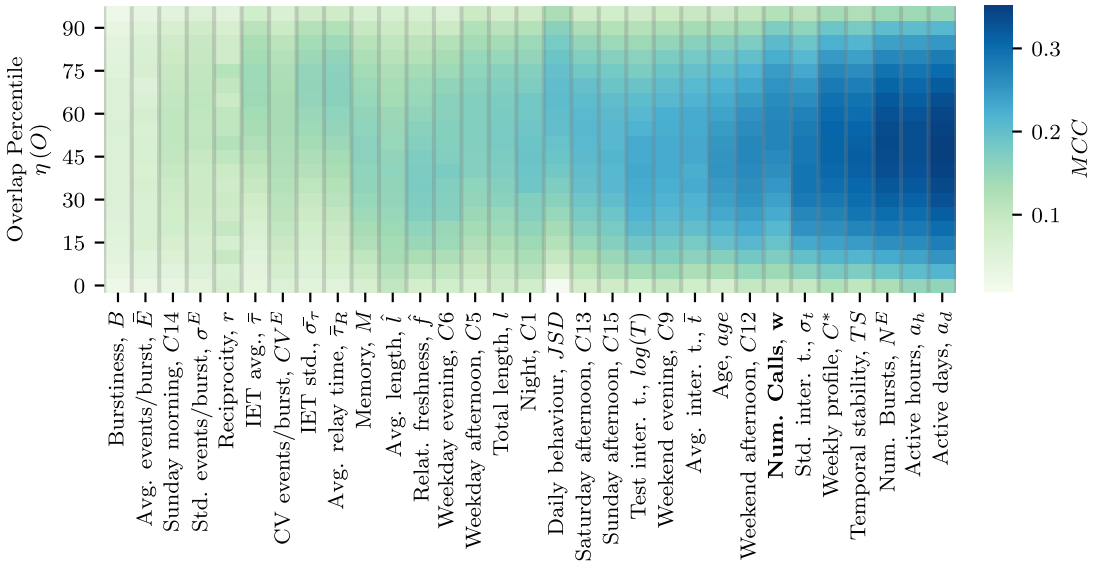

Figure S5. Matthew's Correlation Coefficient (MCC) for overlap prediction with balanced training data, where the x-axis represents variables, the y-axis represents different cutoff values for binary classification, and the color represents MCC. Variables are ranked by their average performance over the overlap distribution.

S2 Predicting Overlap with Balanced Training Data

In the main text, we focus on predicting binary overlap for different high/low cutoff values. We take a naive approach to model fitting in the sense that we do not assume additional conditions on the training data, and as a result or training data is unbalanced for high/low overlap values. We argue that this fitting mechanism provides a clear impression of the range of values where our features are naturally associated to overlap across the whole population. We can, however, improve the performance of our models by balancing our training data via down-sampling, so that in each case the number of high/low overlap is equal. Figure 5 depicts our results using down-sampling.

The main differences concerning the results from the main text is the performance of our estimator in extreme values - MCC is now more smooth along the overlap range. The variable rankings is impacted, although not in a very significant way: our baseline feature w is now the 7th best performing feature. It is worth noting, however, that the MCC values between features in this new ranking does not differ substantially, suggesting that these variables do a similarly good job at predicting overlap.

S3 Predicting Static Overlap

The main body of the text focused on predicting a dynamic measure of overlap on one-month aggregation windows, $\hat{O}^t$. Here we present results where the predictive variable is overlap for the full four-month aggregation window, which we refer to as static overlap O. The conditions of the experiment were identical to those with $\hat{O}^t$: we used the same random sample of 500,000 ties, and predicted different high/low overlap values. We used 3-fold cross validation with four machine learning models ABC, RF, LC and QDA in three scenarios: using each feature as a predictor; using (F, X) variable combinations as predictors, where $F = a_d, a_h, N^E, w$ are the three best-performing variables along with communication

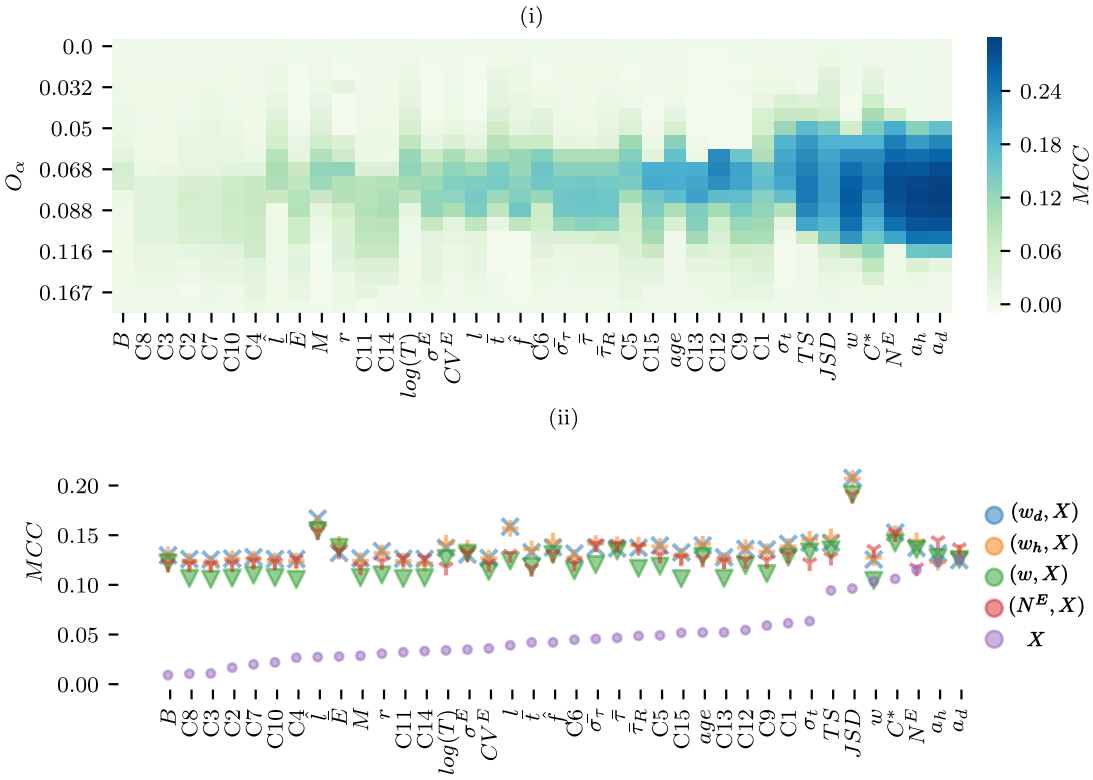

Figure S6. Matthew's Correlation Coefficient (MCC) for static overlap prediction, where the x-axis represents variables, the y-axis represents different cutoff values α for binary classification of high/low overlap, and the color represents MCC. Variables are ranked by their average performance over all cutoff values α. (i) Average MCC for four models trained with single-feature predictors, where each variable is used to predict static overlap using RF, ABC, LG and QDA (ii) Comparison between single and dual-variable models, where we depict the average across all models.

intensity, and X are all the features; and using all the features (the static overlap MCC scores were already presented in Figure 6 of the main text to serve as a point of reference).

Figure 6 our results of static overlap prediction in the single and dual-variable scenarios. The results are relatively consistent to the analysis of overlap prediction in the dynamic scenario, where the number of active days and hours, and the number of bursty trains are the best-performing variables. Although specific variable rankings differ, most variable's behaviour relative to α seem to be consistent (for instance, C^* covers a large range of values, while σ_t's performance is skewed towards lower values). The average performance of dual-variable models differs, however, with the single combination (F, JSD) notably dominating all the other interactions, with a large difference in average performance with the following two combinations, $(F, \hat{l})$ and (F, l), the average and total call length. Compared to dynamic overlap, there is much less variability in combinations (F, X), both regarding F (given a feature X on the x-axis, variables perform similarly) and regarding X (given a top-performing feature F, most variables X perform similarly, with the exception of $X = JSD$).

S4 Overlap prediction using the full set of features

We used the full set of features in the overlap prediction task, with the aim of obtaining maximal predictive performance and understating the relative feature importance (FI), defined as the mean decrease in impurity induced by a feature [2]. Figure 7 displays the maximum MCC for static and dynamic overlap using different models. Both cases follow a similar trend where the best predictive performance is achieved roughly in the middle of the distribution, and where the ML models RF, ABC and LR achieve similar results. Notably, all models perform slightly better for the averaged overlap $\hat{O}_t$, with maximum MCC of 0.457, as opposed to static overlap O, with maximum MCC 0.399.

This can be an indication of the averaged overlap $\hat{O}_t$ being a better proxy for the latent tie strength as discussed earlier. The performance of model QDA is noticeably different from the other three models, outperforming all models for extreme overlap cutoff values but displaying a notably flatter performance curve. As for feature importance, the effect of a_d and $\hat{l}$ dominates other variables in our full model, which is characterized as having skewed FI values. This suggests that a high-performing model can be achieved with a subset of variables, or that some of these variables might contain redundant information on network topology. Many of the variables with high FI, however, correspond to widely different modelling approaches (AP, I, DBC, TS, C_i), suggesting that the interaction between network topology and behavioural features is multifaceted.

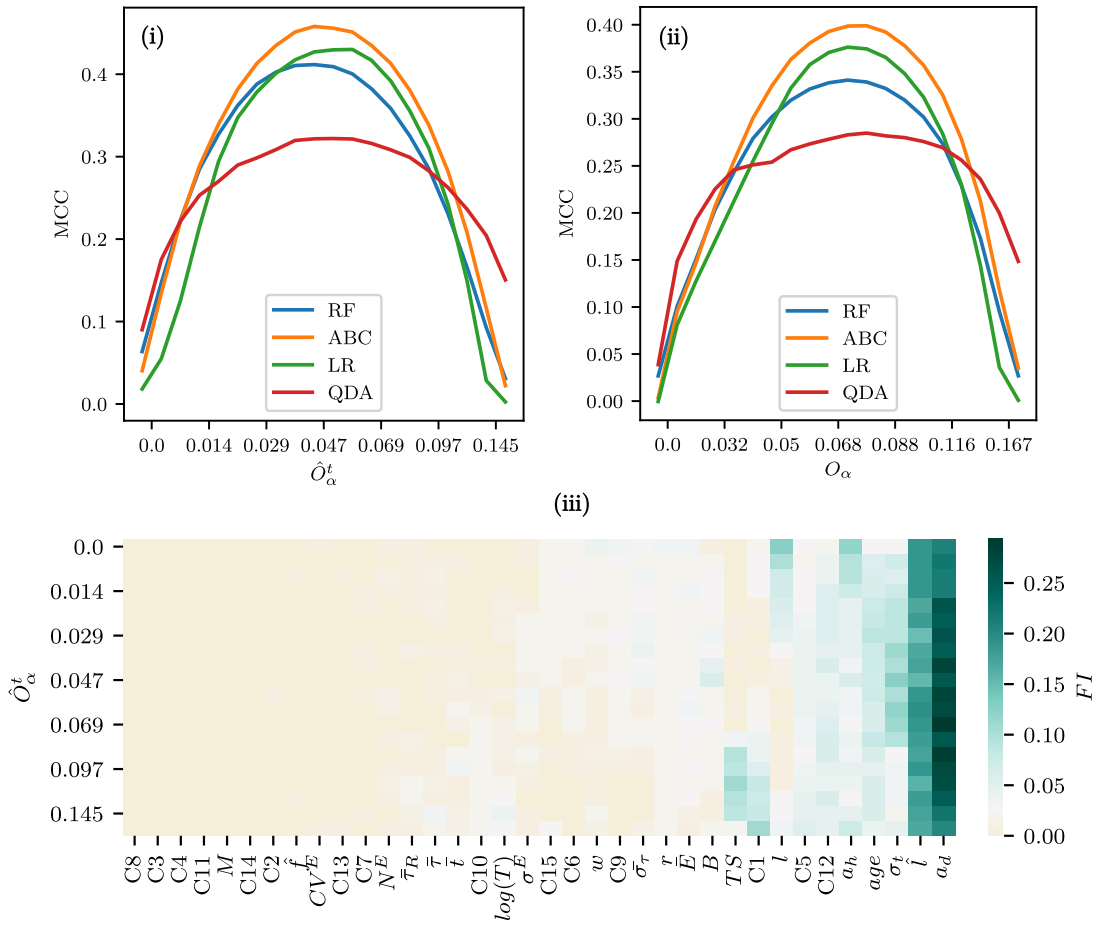

Figure S7. Full model scenario. (*top*) MCC for four different models used in prediction of (*i*) dynamic and (*ii*) static overlap. (*iii*) Feature importance (FI) for the overall best performing model, ABC, for prediction of dynamic overlap. Features are ranked by their average importance over all cutoff values α.

S5 Results by Machine Learning Model

We used four ML models to predict embeddedness using temporal features of human communication, and present the average behaviour of the four models to rank variable performance. Feature performance, however, varied according to the ML models and overlap cutoff values. This can be expected as different ML models focus on different data aspects and might assume different data distributions. In this section we discuss how models affected variable rankings and overall performance.

While a_d remains the highest-ranked feature, there is a higher feature turnover between a_h, N^E and TS, with w now ranked between the fifth and eight position depending on the model. Variables that were not highly ranked for the static scenario are now more prominent, such as age and σ_t, which occurs consistently in all four models. The considerably large range of predictive values of C^* in ABC and RF is now diminished and mostly on par with a_h, a_d and N^E.

Our feature of differences in daily behaviour JSD also ranks consistently lower in this scenario. As with the static case, it performs poorly with RF. For weekly signature clusters, we find a higher prominence in non-linear models RF and ABC. With mean temporal overlap, however, feature $C12$ is now the highest-ranked cluster in all but one model, having a similar behaviour of a small range with more intense predictive value.

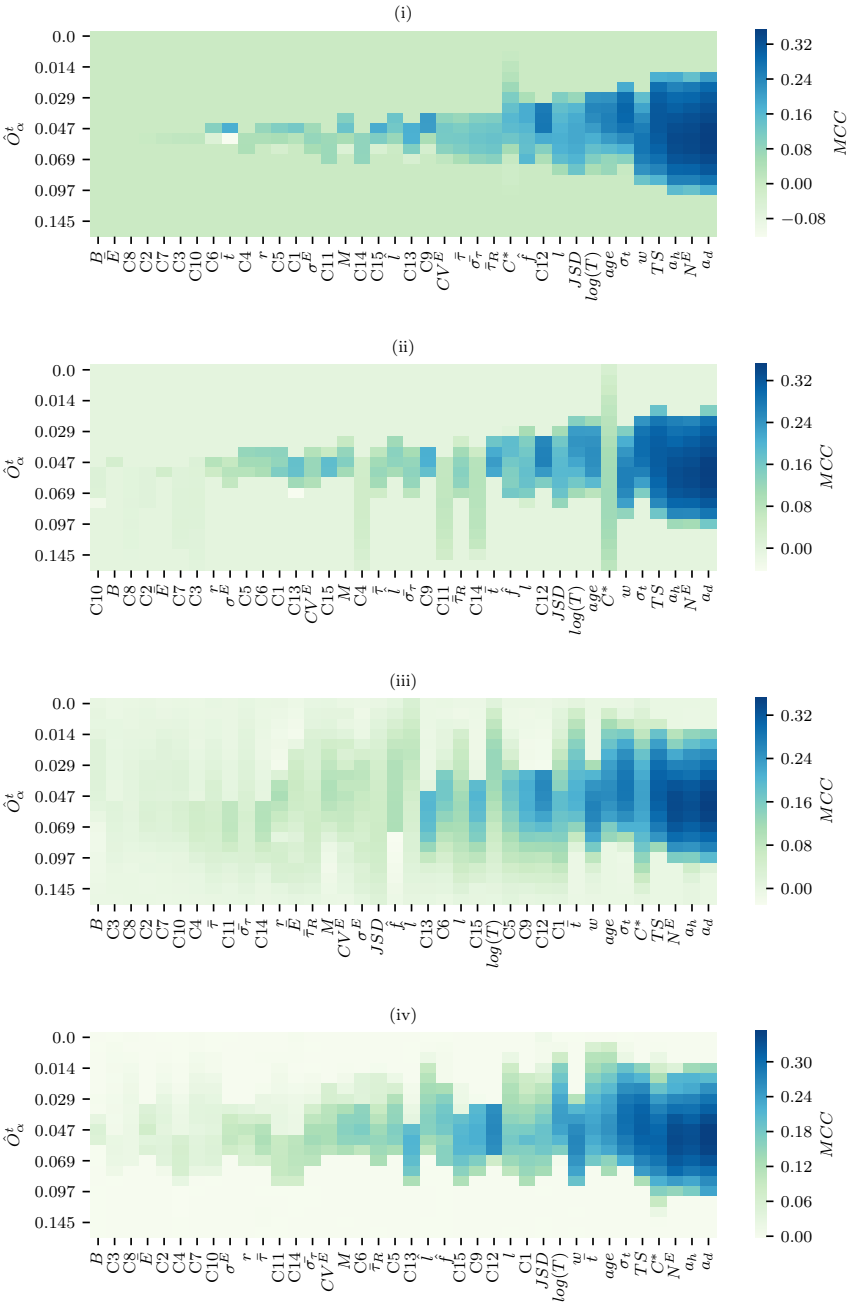

Figure S8. Feature performance in averaged overlap prediction for four different ML models ; (i) LR, (ii) QDA, (iii) RF and (iv) ABC. X-axes represent features ranked by average predictive performance, y-axes represent overlap cutuff values O_α and color represents model performance as measured by MCC.

S6 Additional Performance Scores

We present F1 and AUC evaluation metrics for our main classification problem - the CDR data with a sample of 500,000 ties and 3-fold cross-validation. Since F1 is more strongly affected both by imbalanced training data an specific data labels (binary high and low overlap), we present scores for overlap at a 50th-percentile cutoff value (that is, the median value of the overlap distribution).

Figure 9 depicts the AUC and F1 scores for the prediction task. The rankings follow the general trend observed in the main text, with a_d, a_h and N^E ranking highly in all cases, while being followed by, C^*, TS and σ_t. Notably, for w the F1 score is on par with the highest-ranking features. These metrics suggest that, while specific rankings might differ slightly, the general result that removing burstiness from counts increases performance, while several distinct features capture as much information about network topology as contact counts.

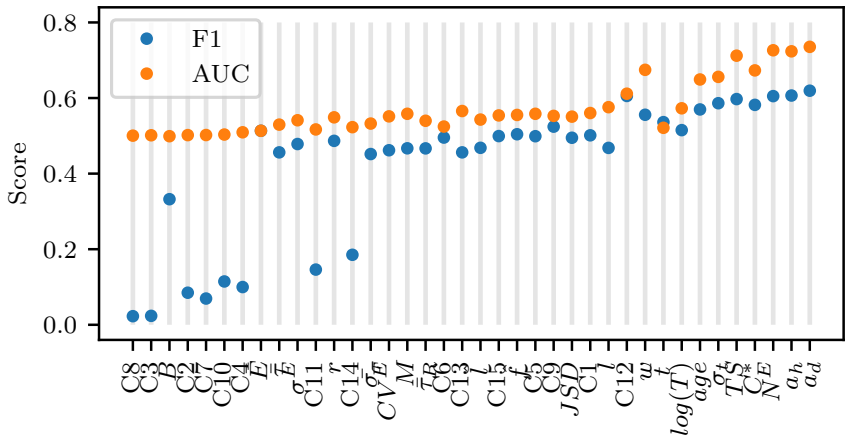

Figure S9. AUC and F1 scores for the main prediction task. Features are sorted according to the ranking of Figure 5 in the main text.

S7 Analysis of bursty cascades

We analyze the effect of the Δt parameter on bursty trains and its relationship to overlap. In order to estimate the effect, we use a sample of 100,000 ties, and use vary Δt on a grid of values with one-hour increments. We evaluate performance via (i) the MCC of a LR that classifies weak/strong ties based on overlap defined at $\alpha = 0.08$, and (ii) Pearson's correlation coefficient with overlap, with results depicted on Figure 10. We choose LR since it is computationally efficient when runnng a large number of cases, and since we know it to perform well for some many of the variables of interest.

Our results suggest some variation both in predictive performance and correlation for different Δt values and for different variables, yet this variation occurs slowly. Prominently, N^E or the number of bursty trains has a predictive performance and correlation that peak at $\Delta t = 26$ and $\Delta t = 32$ hours respectively. We note two particular effects: the dependence of Δt on daily 24-hour cycles, and the interplay between decreasing and increasing trends for different types of variables. In the first case, we note that N^E varies on 24-hour cycles, which is likely due to how Δt overriding the effect of daily

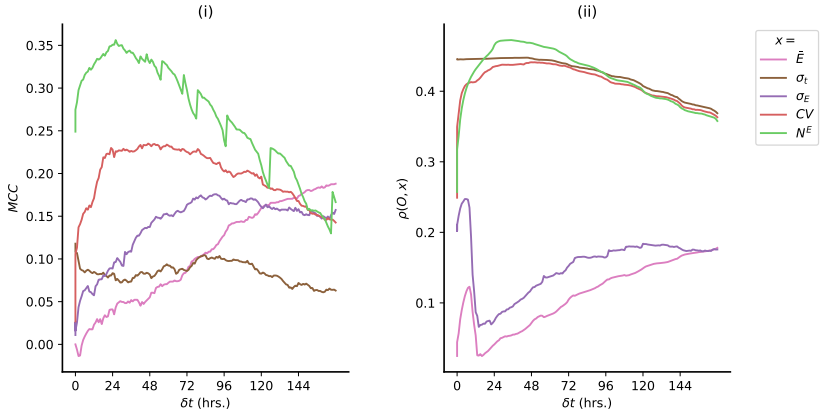

Figure S10. Effect of Δt on the relationship between overlap and variables derived from bursty trains (x). (i) MCC score for LR model for high/low overlap defined at $\alpha = 0.08$ (ii) Pearson correlation coefficient between variables x and overlap.

call-placement cycles. Second, N^E and CV display an initial increase followed by a slower decrease in performance, which constrats to the slower increasing trends of $\bar{E}$ and σ_E. Indeed, since the number of calls w remains constant, for large Δt values the number of cascades decreases, while the number of calls in a cascade E increases, in a manner transferring the information from one variable to the other.

S7.1 Variable Correlations

Figure 11 displays the correlations between the N^E defined for different Δt values. Roughly, the two matrices display high positive correlation, particularly around the diagonal, yet we find three key aspects worth discussing, mainly related to Pearson's correlation matrix.

First, N^E is more sensitive to smaller Δt, where $\Delta t \leq 1$ hour and $\Delta t \leq 1$ day characterize values that correlate the least to the rest of the matrix. That is, the higher-resolution Δt values corrects for the most bursty behaviour. Indeed, since most links have a relatively small number of calls (with higher IET times), the number of bursty trains changes at a slower rate. Second, the matrix roughly follows a block structure, with high correlation on blocks around the diagonal, but where roughly 1-day blocks determine higher or lower correlation, which means that circadian patterns determine the correlation blocks. Last, these differences are decidedly less pronounced for Spearman's correlation coefficient, which implies that the ranking generated by N^E at different Δt values is roughly the same.

S7.2 Number of Bursty Trains

We examine the effect of Δt on the predictive capacity of N^E for different cutoff values of both static and dynamic overlap. The predictive capacity of N^E is greater at around $\Delta t = 1$ day, and does not vary greatly on small changes of Δt. The differences in performance seem to be greater for large differences in Δt.

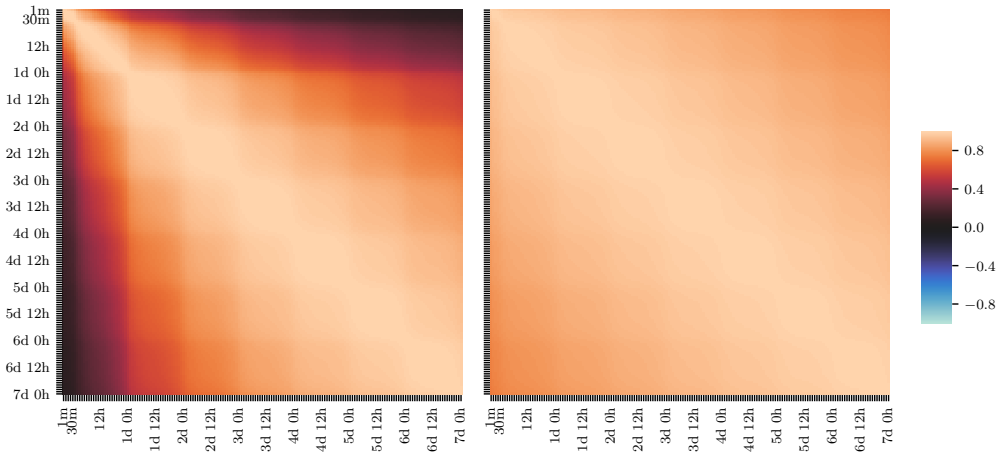

Figure S11. Correlations for number of bursty trains defined obtained for different Δt values; (*left*) Pearson's and (*right*) Spearman's correlation coefficients. We explore Δt values at $\Delta t = 1, 3, 5, 10, 15, 30, 4560$ minutes initially, followed by 1 hour increases.

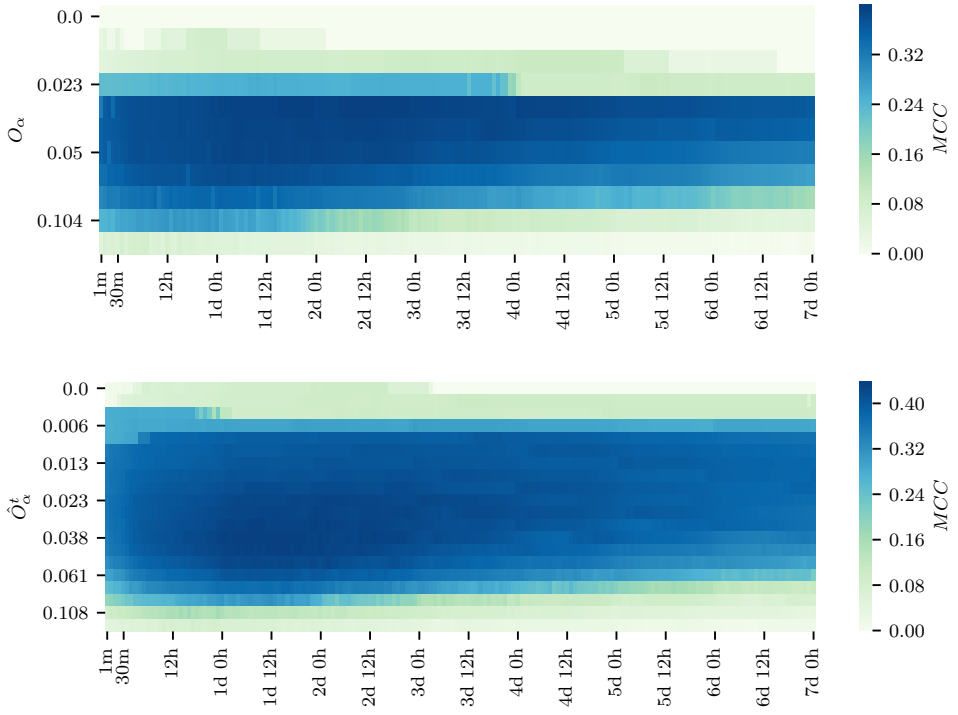

Figure S12. Effect of Δt for using N^E in the prediction of (*top*) static overlap and (*bottom*) mean temporal overlap. The x-axes represent Δt values used to obtain N^E, the y-axes represent static overlap cutoff values O_α, and color represents MCC

S8 Weekly Signatures for Overlap Prediction

In this section, we examine the use of our clustered weekly signatures for overlap prediction. We fit predictive ML models to static and dynamic overlap using only the set of cluster features. Except for the input features, the conditions are the same as in the main paper: we predict binary low/high overlap for different cutoff values, using four ML models (RF, ABC, LR, QDA). We use a random sample of 500,000 ties and 3-fold cross validation to obtain scores and feature importance for the models where it is available.

Figure 13 depicts the MCC scores for each model at each cutoff value, as well as the feature importance of the best performing models (ABC for both cases). These weekly signature models achieve a performance similar to the best-performing variables in the other scenarios. Notably, there are strong performance differences per model, where LR yields an overall poor predictive capacity with a large drop when α is larger. Both ABC and RF capture higher-degree nonlinearities, which together with the poor LR performance could point towards non-linear relationships in the weekly clusters. As with the previous case, our features are able to predict dynamic overlap with a slightly better capacity than the static case (MCC of 0.31 and 0.276, respectively).

Feature importance points towards clusters previously identified as being predictive of overlap for our dataset. The most relevant clusters are $C1$, $C5$ and $C12$, which roughly correspond to late night, weekday worktimes and weekend afternoon, respectively, a result in line with the individual feature performance from before. Our results suggest a nontrivial relationship between the timings of communication events and network topology.

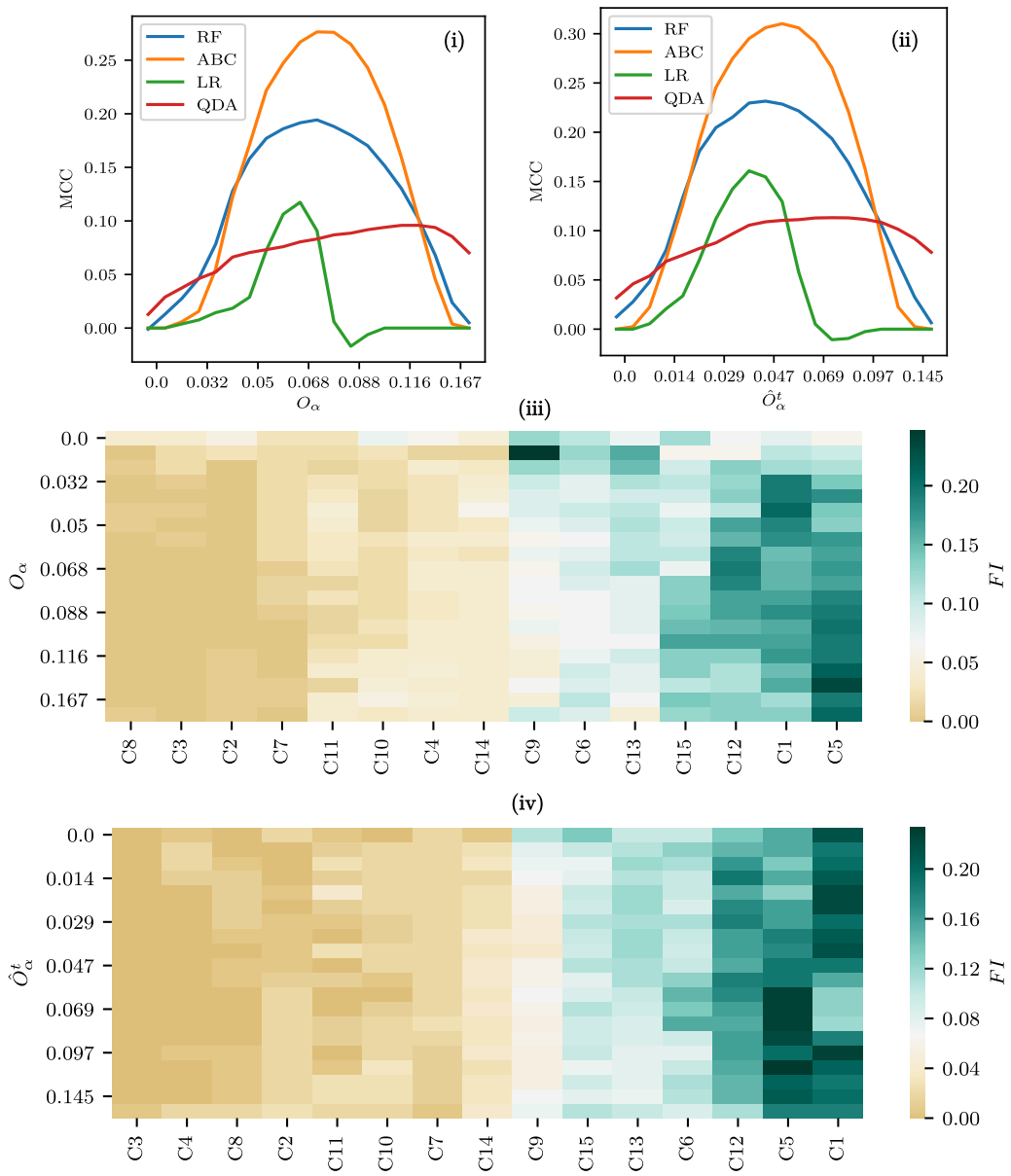

Figure S13. Full model scores and feature importance for weekly signatures. (*top*) MCC for four different models used in prediction of (*i*) static and (*ii*) dynamic overlap. (*iii* - *iv*) Feature importance (FI) for the overall best performing model, ABC, for prediction of (*iii*) static and (*iv*) dynamic overlap. Features are ranked by their average performance over all cutoff values α.

References

[1] Van Dongen, S.: A new cluster algorithm for graphs. Inf. Syst. **1** (2002). doi:10.1046/j.1365-2575.2000.010001001.x

[2] Louppe, G., Wehenkel, L., Sutera, A., Geurts, P.: Understanding variable importances in forests of randomized trees. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 431–439. Curran Associates, Inc., ??? (2013)