communications
physics

ARTICLE

https://doi.org/10.1038/s42005-021-00527-1 OPEN

Emergence and evolution of social networks through exploration of the Adjacent Possible space

Enrico Ubaldi [1], Raffaella Burioni [2], Vittorio Loreto [1,3,4] & Francesca Tria [3,4]✉

The interactions among human beings represent the backbone of our societies. How people establish new connections and allocate their social interactions among them can reveal a lot of our social organisation. We leverage on a recent mathematical formalisation of the Adjacent Possible space to propose a microscopic model accounting for the growth and dynamics of social networks. At the individual's level, our model correctly reproduces the rate at which people acquire new acquaintances as well as how they allocate their interactions among existing edges. On the macroscopic side, the model reproduces the key topological and dynamical features of social networks: the broad distribution of degree and activities, the average clustering coefficient and the community structure. The theory is born out in three diverse real-world social networks: the network of mentions between Twitter users, the network of co-authorship of the American Physical Society journals, and a mobile-phone-calls network.

[1] Sony Computer Science Laboratories, Paris, France. [2] Department of Mathematics, Physics and Computer Science, University of Parma, and INFN, Gruppo Collegato di Parma, Parma, Italy. [3] Physics Department, Sapienza University of Rome, Rome, Italy. [4] Complexity Science Hub Vienna, Vienna, Austria. ✉email: francesca.tria@uniroma1.it

Interactions among individuals shape how our societies unfold and the graph of such interactions can reveal a lot about our social organisation and its evolution in time. That is why social networks have attracted a great deal of attention to understand the mechanisms underlying their evolution and provide valuable information on the microscopic determinants of social dynamics, for instance, individuals' search strategies[1,2] or the schemes to allocate time in socially charged activities[3,4].

The evolution of social networks is shaped by the interplay of complex mechanisms operating at different scales. Indeed, individuals have a heterogeneous propensity to engage in social interactions, featuring heavy-tailed distributions of activity and degree[5]. Also, people allocate their social interactions toward similar alters[6–8], for instance connecting to a friend of a friend-triadic closure[9]. At the same time, individuals may seek novel connections outside of their inner circle of contacts, based on shared interests or experiences (focal closure)[4,10–13]. Moreover, social networks are intrinsically dynamical systems that evolve in time[14,15] as links between nodes are continuously created and destroyed[16–18]. This time-varying nature of the networks deeply affects not only their topological properties[7,15,19] but also the dynamical processes unfolding on their evolving topology[20–23].

The growing availability of large scale and longitudinal datasets logging human interactions allowed for the study and characterisation of the birth and evolution of social networks. This, in turn, triggered the introduction of models capturing some relevant aspects of the whole phenomenology, such as the propensity of individuals to engage in social interactions[5], the correlations in the nodes' activity patterns[21,24,25], the emergence of topological correlations[9,26,27], and the clustering of nodes in tightly connected communities[11,28,29].

All of the models proposed so far, however, feature two main drawbacks. First, most of these models are growing models of a network, and thus they do not account for the dynamics of the network itself. Indeed, the growth is usually simulated by inserting one node per evolution step that establishes edges either by copying neighbours of a randomly selected nodes[13], by rewiring existing connections[30,31] or by following a topological[9,26,27], information-based[32] or hierarchical[33] preferential attachment. Moreover, all of these models return binary networks where no weight is assigned to edges. An exception is given in[34] where, however, authors rely on an initial community structure and degree distribution to reproduce the final modularity of the networks.

The second limitation applies to models describing the network dynamics, allowing nodes to interact more than once in time. Some of them require to fix some data-driven heterogeneous distribution to reproduce the real-world heterogeneity of given observables, such as the fitness associated with each node[9,26] or the propensity for a node to engage in social interactions[28]. Other models focus on particular aspects of the network evolution, such as the drivers of the users' activation patterns in time[35] or the strategies that a node follows to select the interacting partner[36].

In this work, we propose to solve these issues leveraging on the notion of *adjacent possible* space[37–39]. Introduced by the biologist Stuart Kauffman in the framework of molecular and biological evolution, the adjacent possible framework posits that space (be it of words, ideas or products) being explored by some agents is partitioned in three regions: (i) the actual, accounting for all the tokens that users already discovered and experienced, (ii) the adjacent possible space, encompassing all the concepts and tokens that are just one step away from what is known and could become actual in the immediate future, and, (iii) non-adjacent possible space, the set of all those things that could become possible at some later stages, conditional to a suitable expansion of the actual. The key ingredient of this framework is that once an individual experiences something new from his adjacent possible space, the new item gets immediately surrounded by fresh concepts previously belonging to the non-adjacent possible space that is now part of the adjacent possible space. Recently, some of us proposed a mathematical formalisation of the notion of the adjacent possible space[40,41] where the space of tokens that can be explored by an agent (represented, for instance, as a Polya's urn[42,43]) grows when he experiences a novelty, i.e., an item of the space being explored never seen before. This modelling framework allows making quantitative predictions of key statistical features of innovation processes in human activities[44] and technological progress[45].

Here, we apply the adjacent possible framework to describe and reproduce the dynamics of social interactions via a model relying on a minimal set of microscopic rules, i.e., defined at the individual level. We start from the intuition that an exploratory process drives the growth of a social network, as individuals expand their circle of acquaintances by exploring a space of social connections. From this perspective, individuals expand their potential network of contacts—i.e., their adjacent possible space—every time they create a new connection.

We show that our model of network growth and dynamics can correctly reproduce many features of social networks at different scales. For instance, the heterogeneous propensity of nodes to engage social interactions (activity), their degree distribution, the way a node decides to allocate a social interaction toward new or already contacted alters (i.e., the individuals' propensity to innovate and create new connections), the dynamical growth of average degree, the local clustering coefficient, the modularity of a network and the overall rate of growth of the number of edges in it.

Results and discussion

Real-world social network datasets. We test the predictions of our theory against three different real-world social networks that have been extensively studied and characterised in the previous works[21,25,28,46] and that represent a good proxy to dynamically measure interactions between people (see "Methods" and Supplementary Note 2 for details): (i) the American physical society (APS) co-authorship network, (ii) the Twitter mention network (TMN), and (iii) the mobile phone network (MPN).

The most renowned feature of these systems is that both the propensity of a user to engage in a social interaction (i.e., the activity a_i of a node i defined as the number of events actively engaged by node i) and the degree k_i (i.e., the number of different neighbours connected to node i) are broadly distributed. The tails of their distributions are usually approximated with a power-law, i.e.,

$$P(a) \propto a^{-(\eta+1)} \quad \text{and} \quad \mathrm{P(k)} \propto \mathrm{k}^{-\mu},$$

as shown in Fig. 1a, b.

These systems are also expanding in time as new nodes and edges keep entering the network. In Fig. 1c we show the growth in intrinsic time t (i.e., the number of recorded events) of the number of edges $A(t)$ in the systems that follows a Heaps' law as

$$A(t) \propto t^{\gamma},$$

where γ is the exponent leading the growth of the number of links (see Supplementary Note 3 for details).

Another feature is that individuals display correlations on their activity. When a node engages a social interaction, it is likely to turn its social activity (e.g., a mention in the TMN) toward a node already contacted in the past rather than toward a randomly selected node in the system. A possible way to quantify this mechanism is to measure the probability $p_i(k \to k+1)$ (in short

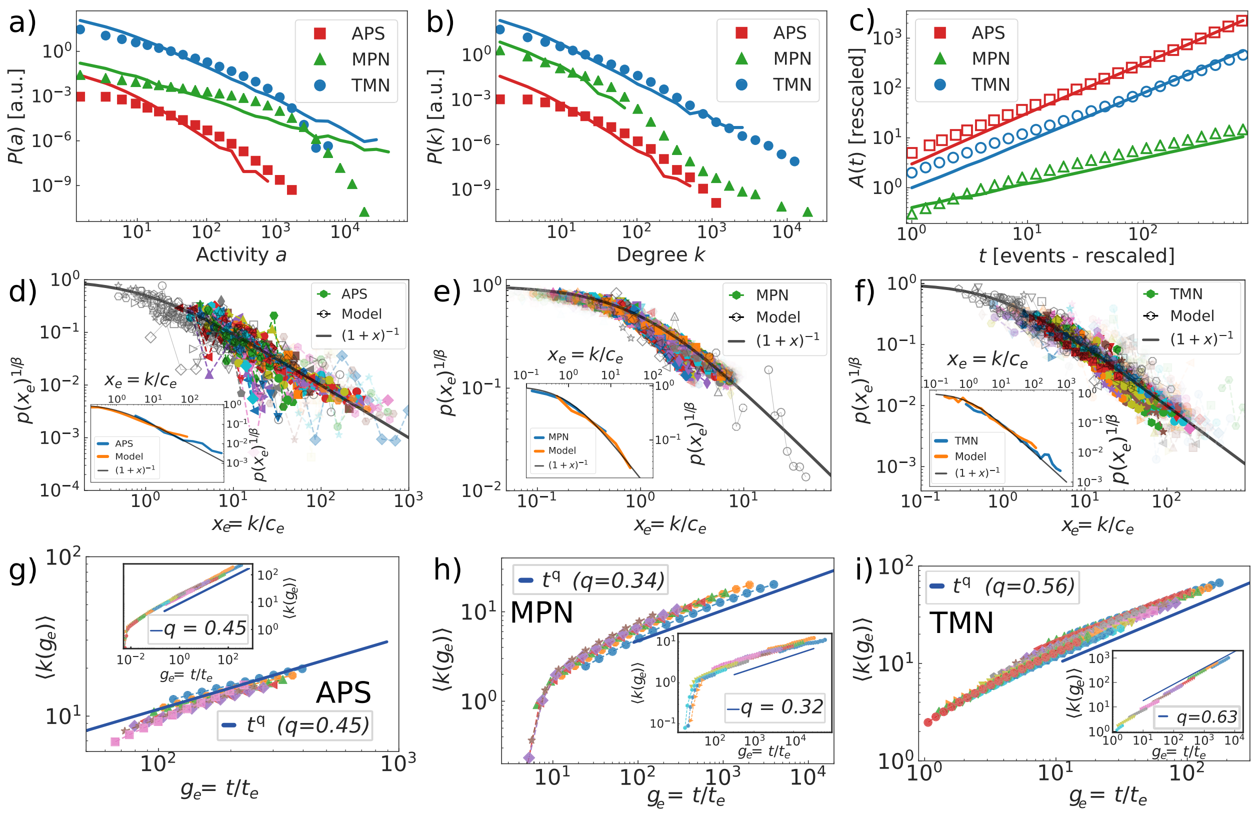

Fig. 1 Stylised facts in social networks' evolution: empirical data and model predictions. a The $P(a)$ activity distribution, (**b**) the $P(k)$ degree distribution, and, (**c**) the temporal growth of the total number of edges $A(t) \propto t^{\gamma}$ as found in the American physical society (APS, red squares), mobile phone network (MPN, green triangles) and Twitter mention network (TMN, blue circles) datasets. In each panel, we also show the same curves as found in the best fitting model of each dataset (solid lines with the same colour as the corresponding dataset). The rescaled strengthening probability $p_e(k) = (1 + k/c_e)^{-\beta}$, being c_e the memory constant for nodes of class e, as measured for different classes of nodes (symbols, colour depth proportional to the number of agents in a node class e) is reported in panel (**d**) for the APS dataset, in (**e**) for the MPN case, and, in (**f**) for the TMN system. In the main panels, we compare the empirical curves (coloured symbols) with the $p_e(k)$ found in the corresponding best-fitting urn model (black symbols) and the theoretical guideline $p_e(x_e)^{1/\beta} = (1 + x_e)^{-1}$ (black solid lines), being x_e the rescaled degree $x_e = k/c_e$. In the insets, we show the average rescaled $\langle p_e(x_e)\rangle_e$ for the empirical (blue lines) and synthetic data (orange lines) as well as the theoretical behaviour $p(x) = (1 + x)^{-1}$ (black line). We fixed c_e by fitting the average $p_e(k)$ as measured for all the nodes belonging to the class e. **g–i** The average degree $\langle k(t_e, t)\rangle \propto (t/t_e)^q$ as a function of the rescaled global time $g_e = t/t_e$ for different classes of nodes entering the system at different times t_e (coloured symbols) for the APS (**g**), MPN (**h**), and TMN system (**i**), respectively. In the insets, we show the corresponding results for the urns model. We also show the best fit $\langle k(t)\rangle \propto t^q$ for all the cases (solid blue lines). Note that here the exponent q does not depend on class e.

$p_i(k)$) for a node i that already contacted k different nodes to contact a new one the next time it will be active[21,25]. The $p_i(k)$, which is formally the probability to pass from degree $k \rightarrow k + 1$, was found to feature the same functional form

$$p_i(k) = \left(1 + \frac{k}{c_i}\right)^{-\beta}$$

across all the analysed datasets[25], with a single value of β -the strengthening exponent- and a distributed, agent-depending strengthening constants c_i (see Supplementary Note 3 for details). At odds with the strengthening exponent β, c_i significantly varies across individuals. To account for the nodes' dynamic variability, we grouped individuals in different classes, e, accordingly to their entrance time t_e into the system and their final degree k_e (see the "Methods" for details). We show in Fig. 1d–f both the rescaled $p_e(x_e)$, with $x_e = k/c_e$ for each class e, (main panels) and the average value of the rescaled probability $\langle p_e(x_e)\rangle_e$ (insets), as found in the empirical and synthetic data (see Supplementary Note 3 for details).

This correlation mechanism inhibits the creation of new links, resulting in a sub-linear growth of the average degree

$$\langle k(t_e, t)\rangle \propto (t/t_e)^q,$$

for the e-th class of nodes, where t is the global intrinsic time measured as the number of total events recorded and where $q < 1$, as shown in Fig. 1g–i.

In Fig. 2 we show for the first time the analysis of Taylor's law[47,48] on social networks evolution data. Recently pointed out as a shared feature in evolving systems, Taylor's law relates the standard deviation of a random variable to its mean and it measures the fluctuations in innovation rate[49,50]. The fingerprint of a complex dynamics is characterised by

$$\sigma(t) \propto \mu(t)^{\delta}$$

with $\delta \geq 1$, at odds with $\delta = 1/2$ characterising uncorrelated events. Complex behaviour of Taylor's law does not trivially follow from Zipf's and Heaps' laws[49,50], making it a relevant observable to test theoretical predictions. We here measure Taylor's law referred to the growth of the individuals' connections

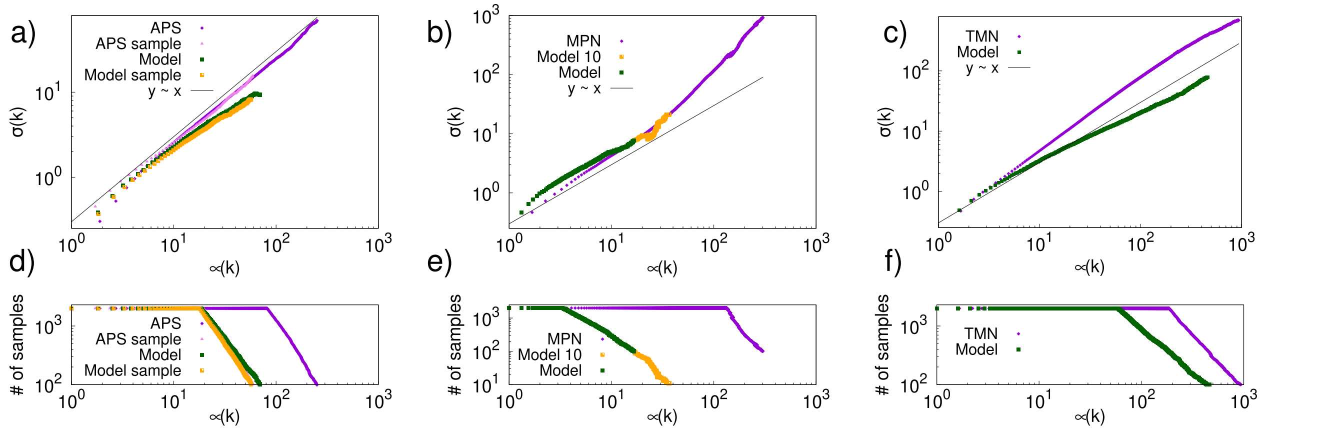

Fig. 2 Taylor's law in social networks. a–c Taylor's law for the number of links created by different individuals at each user's intrinsic time u, being u the number of events where a user took part. We defined $\mu(k) \equiv \langle k_i(u) \rangle_i$ and $\sigma(k) \equiv \sqrt{\langle k_i^2(u) \rangle - \langle k_i(u) \rangle^2}$, where $k_i(u)$ is the degree of node i at its intrinsic time u and where we implicitly have a dependence on the node i's intrinsic time u. Results in the empirical datasets, contrasted with model predictions, for (**a**) the American physical society dataset (APS), (**b**) the mobile phone network (MPN) dataset, and (**c**) the Twitter mention network (TMN) dataset. In (**a**) we also report results for the 1-link sampled database, defined in the "Methods" section. For all the empirical and synthetic datasets we find an approximate power-law behaviour $\sigma(k) \sim \mu(k)^\delta$, with $\delta \gtrsim 1$. For each curve, we considered the 2000 individuals with the longest history (final intrinsic time). **d–f** We report the effective number of points over which the mean and the standard deviation is computed at each time t. This number decreases for the highest values of the mean, corresponding to higher values of intrinsic time since the number of the total performed actions varies from individual to individual. The number of samples used is reported, both for the empirical and synthetic datasets, for (**d**) the APS dataset, (**e**) the MPN dataset, and (**f**) the TMN dataset. In all the cases, we considered only averages on at least 100 samples, but for the MPN case, where we additionally show values for times where at least 10 samples were available (model 10 curves).

in the network, observing a linear or superlinear behaviour (Fig. 2).

Besides these fundamental features, we also track a set of local and global observables, later presented in the Results section.

Model. As said, we build on the adjacent possible framework[40] to model the exploration of social spaces where individuals are embedded. Within this framework, we microscopically model how the space of possibilities of a node (i.e., the set of all the social interactions that are "possible" for a node) evolves in time. This space, at a given point in time, consists of three distinct regions: (i) the *actual*, including all the links already experienced by the individuals in the past (current connections), (ii) the adjacent *possible* space accounting for all the links that are just one step away from being explored (e.g., the friends of friends not yet our friends), and, (iii) the *non-adjacent possible* space, accounting for all the links that may become adjacent and possible at some later stage.

A second essential ingredient of our model is the presence of correlated novelties[40]. Every time the social exploration process of a node i activates a new connection with a node j belonging to its adjacent possible space, i and j experience a novelty, i.e., the link e_{ij} gets active for the first time. In this way, j becomes now part of the *actual* region of i and the *adjacent possible* space of i enlarges with new possible connections that were not possible for i before. In other words, a novelty paves the way to another in the future.

The modelling scheme we apply is a multi-agent version of a modified Polya's urn[42,43] that has already been successfully implemented to reproduce the key statistical properties of complex systems (Zipf's, Heaps' and Taylor's law) and their dynamical correlations[40,44,50,51]. In the simpler formulation of that model[40], the key ingredient is an urn, $\mathcal{U}$, initially containing N_0 distinct elements. One may think of them as balls of different colours representing an item of the space being explored. The dynamics proceed by repeatedly withdraw balls from $\mathcal{U}$ and annotating them in a temporal sequence of events $\mathcal{S}$-here this sequence represents a sequence of social contacts experienced by a user. Every time we pick up a ball, we put it back in the urn together with ρ additional copies of it, thereby reinforcing that element's likelihood of being drawn again in the future, in a "rich-get-richer" fashion. To account for the adjacent possible space expansion, whenever a novel (never extracted before) element appears in the sequence $\mathcal{S}$, we additionally put $\nu + 1$ new distinct elements in $\mathcal{U}$, thus expanding the adjacent possible space of the system.

We generalise this model to a multi-agent definition to account for the birth, evolution and dynamics of social networks. To this end, we introduce two key concepts. First, the system consists of a collection of urns, each identified by a unique alphanumeric ID (a, b, c,...), representing users in a social network. Second, each ball within each urn bears the reference ID of another urn in the system. Then, the sequence of extracted balls will correspond to a series of social contacts annotated as tuples (i, j), where i is the ID of the urn drawing a ball, and j is the ID of the drawn ball. For each extraction, the reinforcement process requires to put back ρ copies of the extracted ball j into the extracting urn i (and vice-versa), so that an exploited interaction will be favoured again in the future. To account for the expansion of the adjacent possible space, we also let two urns that interact for the first time to exchange a *memory buffer*, i.e., a subset of $\nu + 1$ balls that each urn shares with the other. This set is selected using a specific rule s that we present below. Thanks to this exchange, an urn that experiences a novel connection expands its adjacent possible space, thus increasing the set of IDs that it may contact in the future.

Exploration strategies. The last ingredient of our model is the strategy s that an agent adopts when sharing its experience (the memory buffer) with nodes contacted for the first time. We introduce a total of six different strategies s to determine the $\nu + 1$ IDs contained in the memory buffer being shared along with new links. Here, we report three of these strategies

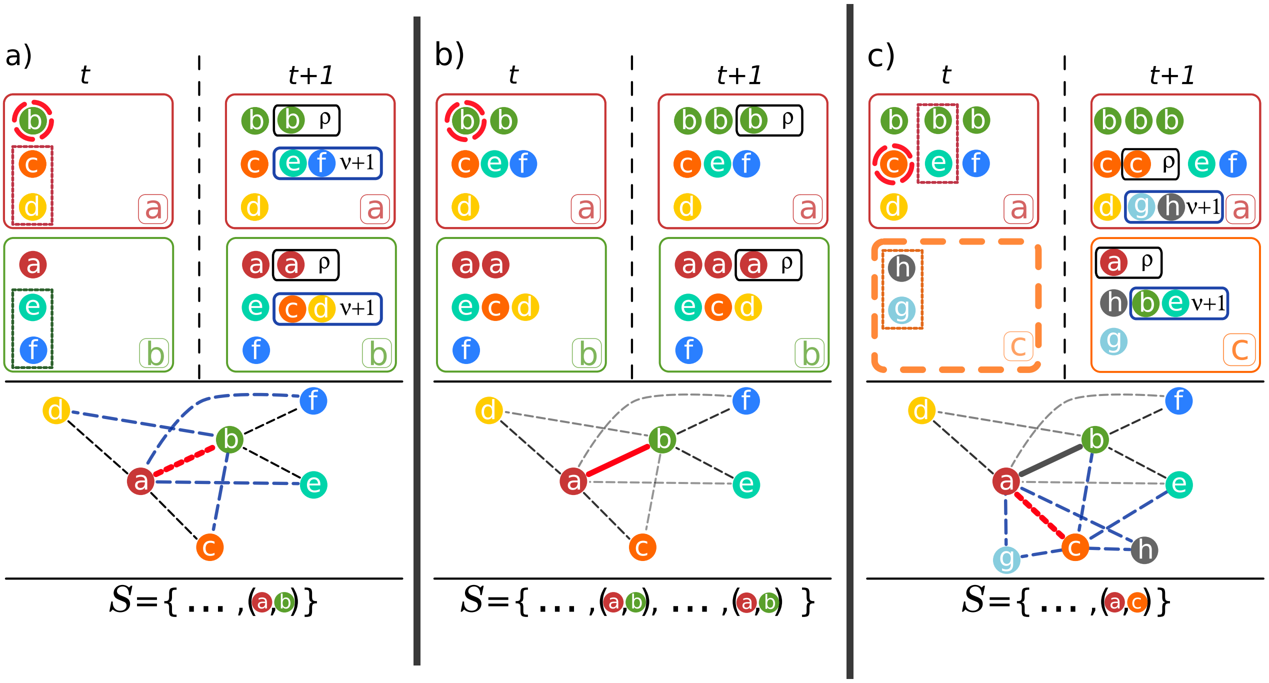

Fig. 3 Three possible evolutionary steps of the model. Three possible steps of the Polya's urn model for a system with equal reinforcement and innovation parameters $\rho = \nu = 1$ with the weighted sampling with withdrawal sampling strategy $s =$ WSW. For each evolutionary step of the system (columns), we show the current state of the urns (top row), the equivalent network evolution (mid-row), and, the sequence $\mathcal{S}$ of observed events (bottom row). In the network, we show already active links (solid lines), links in the adjacent possible space (dashed lines), currently active links (red lines) and connections entering into the adjacent possible space (blue dashed lines). The ball drawn is visualised with a red dashed circle while we show reinforcement balls and memory buffer ones enclosed in black and blue rectangles, respectively. New urns entering the system are shown with dashed borders. **a** At time t urn a is active and draws the ball b: the event (a, b) is then appended to the sequence $\mathcal{S}$. At time $t+1$ the urn a then gains ρ copies of b and vice-versa (reinforcement) and, since the e_{ab} link is new, we also draw $\nu+1$ distinct balls from a following the WSW strategy (balls c and d within the dashed rectangle) that will be copied into b (and the same for b that sends e and f as novelties to a). In the network representation, the e_{ab} edge is active and the e_{ae}, e_{af}, e_{bc}, and e_{bd} links enter into the adjacent possible space. Notice that the adjacent possible of c changed without the need for c to participate in social interaction. **b** At time t, urn a draws a copy of b (top). Since the edge e_{ab} was already active in the past, we only put ρ copies of b in urn a and the other way around. The network's topology does not change in this step, while the weight of the e_{ab} link gets increased (network representation). **c** Urn a draws a copy of c at time t (top). Since c is an empty urn, it creates $\nu+1$ novel IDs (g and h, in the dashed rectangle) and gains a copy of them. We add (a, c) to the sequence $\mathcal{S}$ and we perform the reinforcement/novelties exchange between a and c. The network gains two new nodes (g and h), activates a new edge (e_{ac}) and inserts new links in the adjacent possible space. The actual space of c acquires a while its adjacent possible space gains e, g, and h.

that turn out to best capture the phenomenology of the empirical datasets we consider, while we refer to the Supplementary Note 1 for the other strategies. The first strategy is the weighted sample with withdrawal (WSW) strategy: an agent draws $\nu+1$ *distinct* IDs from the urn proportionally to their abundance in the urn itself at that time, i.e., proportional to the number of the past interactions with each ID. Distinct means that the node extracts exactly $\nu+1$ IDs proportionally to their abundance and withdraws all the balls of an ID x after it has been drawn. This strategy corresponds to sharing the IDs that interacted the most with a node in the past and is the one applied in Fig. 3. Then, we define the symmetric sliding window (SSW): each agent keeps a buffer of its last $\nu+1$ distinct alters with whom he interacted in the past. These represent the set of IDs shared with a newly contacted agent. After the exchange, both agents update their memory buffer by pushing in the ID of the agent just contacted and removing the $\nu+$ 1st ID from their buffers. This strategy favours the spreading in the network of the recently activated connections, rather than the most frequent ones. Finally, we introduce the asymmetric sliding window (ASW): it is a variant of the previous strategy, the difference being that only the agent that initiated the interaction updates its memory buffer after the communication event. Given these definitions, we can now define the sequential evolution steps of our model.

Model rules. The three parameters of the model are the reinforcement parameter ρ, the number of novelties to be shared ν and the memory buffer exchange rule s. A schematic representation of the model is given in Fig. 3 and we resume here the steps defining it (see the Supplementary Note 1 for details):

(1) we start with two urns, a and b having a copy of each other's ID inside of them; the urns also contain the $\nu+1$ distinct identities (IDs) of other urns that did not participate yet to any interaction (c, d for a and e, f for b). These sets represent their initial *memory buffers*. The sequence of events $\mathcal{S}$ is initially empty;

(2) at each time step, we extract a "calling" urn i proportionally to the size of the urn U_i (the number of balls within the urn i). We then draw a ball from the calling urn i, say the ID j. This double extraction corresponds to a single event (i, j) that we append to the main sequence $\mathcal{S}$. In Fig. 3 the first event is the (a, b) one.

(3) reinforcement: following the event (i, j), we add ρ copies of i in the j's urn and ρ copies of j in the i's urn. For example, in

Fig. 3a, b, we add ρ copies of a in the b's urn and ρ copies of b in the a's urn.

(4) novelty: if it is the first time that i and j interact, i and j exchange their *memory buffer*. With this mechanism, we add j's memory buffer into U_i and, vice-versa, i's memory buffer into U_j. In Fig. 3a, a's memory buffer (c, d) is copied into U_b and b's memory buffer (e, f) is copied into U_a. In this case, the memory buffer is determined using the $s = WSW$ strategy, that is, by extracting exactly $\nu + 1$ IDs from U_a proportionally to their abundance.

(5) if a node j is called for the first time by another node (i.e., j is an empty urn so that $U_j = 0$), it creates $\nu + 1$ new agents (empty urns) and, for each of them, it creates a ball into its urn: these $\nu + 1$ IDs represent the initial memory buffer of j. In Fig. 3c node c creates two brand new nodes, g and h, that will represent its initial memory buffer. We note here that the newly created agents are initially empty urns so that they can participate in the dynamics (they can be included in the social network) only if another urn (agent) calls them. Only after this first call, they may actively engage in an interaction. In this scheme, an agent cannot join the network "from outside," i.e., unless it is engaged by another agent already belonging to the network. Of course, the scheme can be generalised to account for other schemes for nodes to enter the system but we will not cover this detail in the present work.

Each evolution step is defined as a repetition of the 2 → 5 steps of the just outlined procedure, as shown in Fig. 3. The parameters ρ and ν weigh the relative importance of the reinforcement and exploration processes in the system. We define $R = \rho/\nu$ as the ratio between the two. The model is then entirely defined by three parameters only: the reinforcement value ρ, the ratio $R = \rho/\nu$ setting the relative importance of the reinforcement (exploit) and novelties (explore) mechanisms, and the strategy s used to exchange the memory buffer between nodes getting in contact for the first time.

Let us note that our model does not only represent a network growth model. Instead, it reproduces a sequence of interaction events that mimic the original network dynamics, i.e., the actual sequence of events. This feature is at variance with other models of network growth where each evolution step corresponds to the insertion of a new node establishing its connections following a given rule[9,13]. Moreover, the creation of new edges and the reinforcement of their weights stem from the exploration that each user does of the possible connections and the expansion of the number of potential acquaintances rather than from local rewiring schemes[30,31]. In the following, we show that, after fitting the model's parameters to the three different networks we analyse, the model's simple microscopic rules reproduce the main topological and dynamical features of empirical networks. Besides, the values of the three model's parameters give some insights on the nature of the three social networks we are analysing.

Comparison between model and datasets. We now compare the main network features emerging from our modelling scheme's evolution to their empirical counterpart. To compare the theoretical predictions with the observed data, we optimised the model by fixing the parameter values that minimise, for each empirical dataset, a cost function $S_d(\rho, R, s)$. The latter evaluates the goodness of fit of the synthetic simulations to the empirical dataset d by looking at eight selected observables, both local and global, topological and dynamical (see the "Methods" for details). Specifically, we consider (i) the strengthening exponent β, (ii) the q exponent leading the average degree growth, (iii) the γ exponent driving the growth of the number of edges growth in time. We also take into account (iv) the asymptotic value of the average clustering coefficient c and the fraction of activated edges that are either old (already activated in the past) or new (being activated now) and that happen to insist on a triangle (closed) or not (open), thus defining four categories: the (v) old open (OO), (vi) the old closed (OC), (vii) the new open (NO) and (viii) the new closed (NC) active links, measuring the fraction of events falling in each category per time range in the asymptotic limit of the system evolution (see the "Methods" section). We summarise the results in radar plots in Fig. 4, showing the observed values for the eight selected observables along with their best numerical estimates. We observe that the model endowed with optimal values of the parameters is able to quantitatively reproduce all the selected observables in all the datasets, exception made for the APS dataset. In the latter, the model fails to predict the observables related to the network topology correctly. To explain this discrepancy, we note that the APS dataset is composed of cliques of events—rather than one single event between two IDs per time. This feature leads to a high clustering coefficient (as all the agents publishing one paper are fully connected) and in an increased count of events observed along old edges insisting on at least one closed triangle. To filter out this effect, we performed a subsampling of the data by drawing a single link among all the possible ones for each paper and re-computed the features of this sub-sampled dataset (see the "Methods" for details and the Supplementary Note 4 for other sampling strategies that give similar results). When analysing these results, the disagreement between the real system and the model results disappears (see Fig. 4a, b), revealing that the model can explain the underlying interaction processes also in this dataset.

In the Supplementary Note 1, we give an analytical solution of the model and we show how each model parameter affects the main observables of the system. Here we report that as for the global observables discussed in "Results and discussion" and reported in Figs. 1 and 2, the relevant parameters are the ratio $R = \rho/\nu$ and the sharing strategy s, while the absolute values of ρ and ν impact the behaviour of the observables related to the local topology of the network. Despite the limited number of parameters, the model is flexible enough to reproduce a wide range of phenomenologies, from highly exploratory situations such as the APS dataset (high γ and low β exponent) to more exploitative scenarios, with a reduced number of connections being explored such as in the MPN (with a large β and a low γ exponent).

Global trends. Besides the eight observables included in the cost function $S_d(\rho, R, s)$, in Fig. 1 we show that the model can also reproduce the empirical networks' broad activity (panel a), the degree distributions (panel b), the growth of the number of edges in the network $E(t) \propto t^\gamma$ (panel c), the functional form of the strengthening function $p_e(k)$ (panels d–f), as well as the sub-linear growth in of the average degree (panels g–i). Let us note that the average $p_e(k)$ curves presented in the insets of Fig. 1d–f feature different support in the rescaled degree $x_e = k/c_e$. In particular, the APS and TMN empirical data span a larger rescaled degree range with respect to the synthetic data, while the opposite is true in the MPN case. This difference is due to the discrepancy in the $P(c_i)$ distribution of the strengthening constant c_i in the empirical populations and the synthetic ones, as we show in the inset of Fig. 5a. Indeed, our model reproduces larger (smaller) c_i in the APS/TMN (MPN) cases compared to the empirical situation. This results in a smaller (larger) range of the rescaled degree (k_e) as shown in Fig. 1d–f.

In the insets of Fig. 1g–i, we also show that the model can reproduce the average degree growth in time $\langle k(t_e, t)\rangle \propto t^q$ for

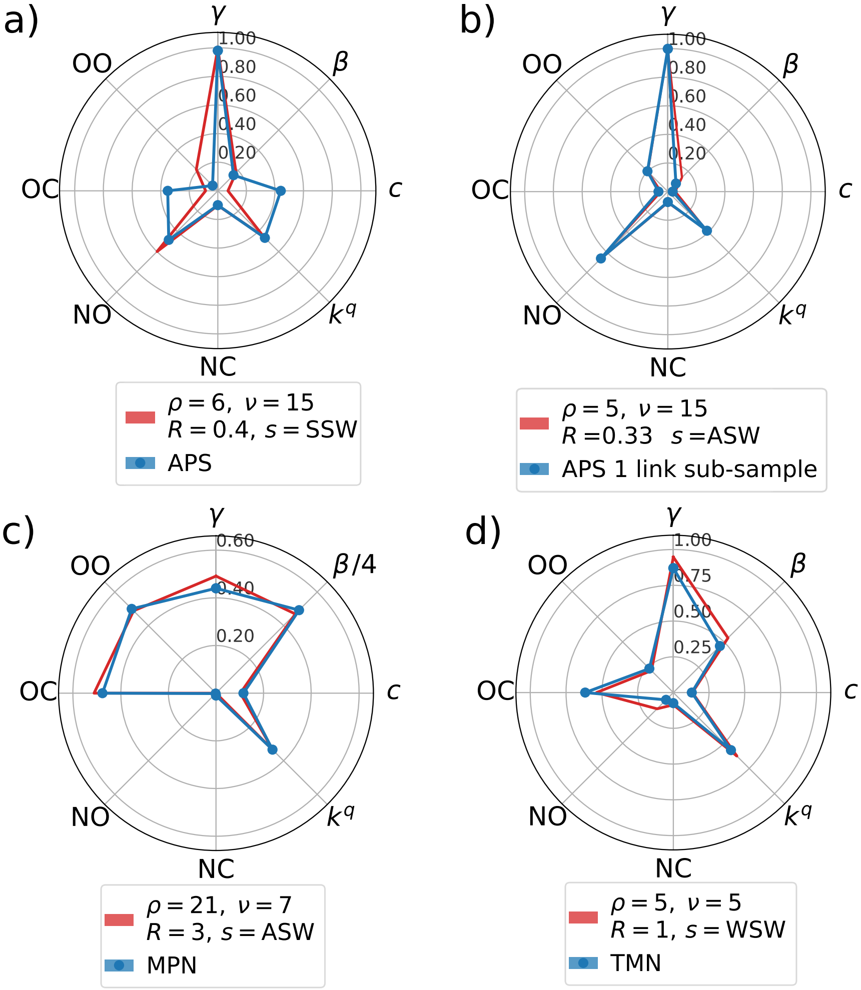

Fig. 4 Eight selected observables for model's optimisation. a–d Radar plots comparing eight selected observables measured in empirical (blue lines) and synthetic data (red lines). In each panel, we report the reinforcement and novelties parameters ρ and ν and their ratio R as well as the optimal sampling strategy s, that is: (**a**) American physical society (APS) with the symmetric sliding window (SSW) strategy s, (**b**) the 1-link subsampled APS with the asymmetric sliding window (ASW) strategy, (**c**) the mobile phone network (MPN) fitted with the ASW strategy, and, (**d**) the Twitter mentions network (TMN) with the weighted sampling with withdrawal (WSW) strategy.

the different node classes. The q exponent reproduced by the model is not always in perfect agreement with the empirical value (e.g., $q = 0.56$ in the TMN while $q = 0.63$ in the best fitting model), but in the Supplementary Fig. 13, we show that the distribution $P(q_e)$ of the exponent per node class is in very good agreement between the empirical and synthetic cases.

Furthermore, the model can reproduce the broad fluctuations in the rate of innovation (i.e., in the creation of new links), as measured by Taylor's law (Fig. 2). In the Supplementary Note 3G we also show that the model qualitatively reproduces the inter-event distribution between two consecutive events of edges creation for a node as well as the entropy of such sequence. These results represent the first improvement of our model with respect to the state of the art as we do not require (i) to manually specify a broad fitness distribution of the nodes to reproduce the heterogeneous degree distribution as in ref. [9], (ii) to specify the probability to contact a new node instead of an already contacted one as in ref. [28], as the strengthening behaviour of $p_e(k)$ naturally emerges from the microscopic evolution rules, and (iii) to set the selection rule of the next active node i and the probability distribution of the alter j that he will contact, as they emerge from the model as opposed to refs. [26,34].

Heterogeneities in the experience of the new. Besides the properties already exposed, the model also correctly captures the heterogeneous propensity of individuals to establish new connections, i.e., the rate at which they experience novelties. To quantify this rate, we look at the exponent of the Heaps' law describing the growth of the degree of an individual, $k_i(x_i)$, i.e., the number of distinct people encountered as a function of the number of social events performed x_i: $k_i(x_i) \propto x_i^{\alpha}$. Figure 5a reports the distribution of empirical exponents α for the three datasets considered. These distributions are peaked at different $\bar{\alpha}_d$ values for the different datasets ($\bar{\alpha}_{\mathrm{APS}} \sim 0.9$, while $\bar{\alpha}_{\mathrm{TMN}} \sim 0.7$ and $\bar{\alpha}_{\mathrm{MPN}} \sim 0.4$). Remarkably, the model correctly reproduces both the peak value and the broadness of each empirical $P(\alpha)$ distribution.

Another empirical quantity heterogeneously distributed is the strengthening constants c_i, setting the probability for an individual with k connections to acquire a new one as $p_i(k) = (1 + k/c_i)^{-\beta}$. The inset of Fig. 5a shows that our model qualitatively reproduces the empirical distribution $P(c_i)$ of the strengthening constants c_i, even tough with some discrepancies between the empirical and synthetic data. In particular, the variance of the synthetic values is limited by the strict evolution rules of the urn, that cannot indefinitely diverge from the average

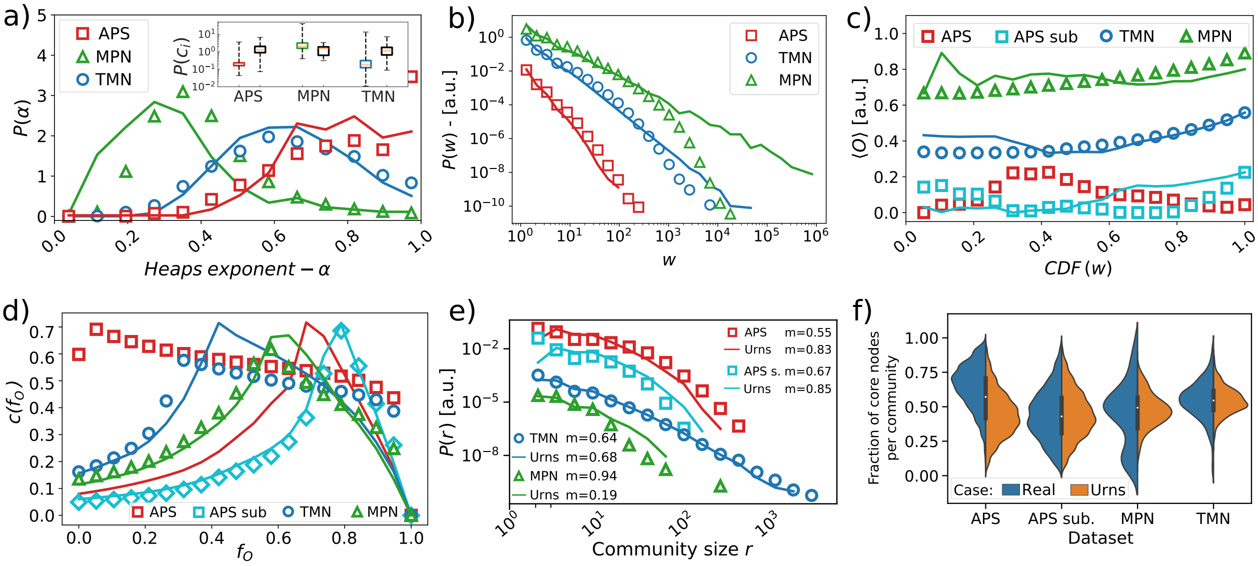

Fig. 5 Topological properties of social networks: empirical data and model predictions. In all the panels, we show the data of the empirical datasets using symbols and data from the synthetic simulations with solid lines. To distinguish the datasets, we use the red colour for American physical society (APS), cyan for the APS 1-link subsample dataset, green for the mobile phone network, and blue for the Twitter mentions network dataset. **a** The $P(\alpha)$ distribution of the local Heaps' exponent α for a sample of 10% of nodes as measured in the empirical networks and the simulated ones. In the inset, we show the $P(c_i)$ distribution of the individuals' strengthening constant c_i as measured in empirical data (coloured boxes) and their corresponding simulations (black boxes). **b** The $P(w)$ link weight distribution for the three empirical datasets (symbols) and the ones found in the artificial networks (solid lines). **c** The average overlap in the network obtained by removing edges in ascending order of weight up to a given cumulative density function value (percentile). **d** The average clustering coefficient $c(f_O)$ measured on the network obtained by removing edges being in the overlap percentiles $O \leq f_O$. **e** The distribution $P(r)$ of the community sizes r found in the different datasets. In the legend, we show the network modularity m as measured in each dataset. **f** The distribution of the fraction of core nodes within the communities of the different networks. We compare the empirical results (blue distributions on the left) with the synthetic ones (orange areas on the right, the dataset names are reported on the x-axis).

behaviour. That is why all the three best-fitting models feature a similar $P(c_i)$ distribution (all peaked at around $c_i \sim 1$) whereas the empirical cases show smaller c_i in TMN and APS, while larger c_i are observed in the MPN case. A possible extension of the model should then try to address this drawback. However, the model automatically reproduces a distributed propensity of individuals to decrease their social exploration at a given cumulative k. This is at variance with previous works that explicitly encoded this rule in the model definition[28,35].

Topological correlations. We now focus on the topological correlations of the empirical and synthetic networks of interactions. In Fig. 5b we show that the model correctly reproduces the overall link weight distribution $P(w_{ij})$, i.e., the distribution of the number of activations of a single edge w_{ij}. Then, we test that both the empirical and the synthetic data obey the weak and strong ties scheme of the Granovetter conjecture[2,12]. The latter states that links in a social system will be arranged to have communities of individuals tightly connected by strong ties and with a large neighbours overlap. These communities are then interacting through weak ties, i.e., links acting as bridges between communities composed by nodes sharing a limited number of common neighbours (low overlap). To prove this, we measure how the overlap O_{ij} of two nodes (i.e., the fraction of common neighbours of nodes i and j with respect to their total number of neighbours) correlates with the weight w_{ij} of the edge between i and j. In Fig. 5c we show how the average edges' overlap varies as we filter out network's edges with a weight smaller than a given percentile w. In all the cases we observe a positive correlation between the two quantities at high values of w, the only exception being the APS case (red symbols). As for the previous clustering analysis, we show that once we subsample the APS events, the empirical data agree with numerical simulations (cyan symbols and line). Also, in this case, our model is able to reproduce a non-trivial arrangement of the edges' weights and their topology. In the Supplementary Note 3H we compare other topological measures finding a good agreement between the empirical and synthetic cases (positive degree assortativity and negative local clustering-degree correlation).

We further test if our model reproduces the community structure (modularity) of the empirical networks. In Fig. 5d, we inspect how the average clustering coefficient $c(f_O)$ of the network varies when removing the edges e_{ij} by their ascending overlap, i.e., by discarding edges with overlap $O_{ij} \leq f_O$, being f_O the cumulative percentile of the overlap distribution. We find $c(f_O)$ to increase as one removes edges with small overlap, indicating that the removal of weak ties is removing bridges between communities. Then, the $c(f_O)$ peaks and, if we keep removing the higher overlap edges, we start breaking the triangles in the communities' cores, the clustering coefficient decreases. Again, the disagreement between the empirical and synthetic APS dataset disappear if we consider the 1-link sub-sampled dataset. As shown in Fig. 5d, the subsampled dataset perfectly matches the corresponding numerical data.

Then, in Fig. 5e we show, for each dataset, the $P(r)$ distribution of the community size r as found in the empirical networks (symbols) and the numerical simulations, together with the modularity values of the networks (see the "Methods" for details). Surprisingly enough, the model can reproduce the communities' size distribution and the modularity values for the APS (both original and sub-sampled) and the TMN datasets. In the MPN dataset, the synthetic network's modularity is found to be smaller than the empirical one. This finding can be due to the limited size that the network reaches in the simulation time (~2500 nodes versus the millions of the real case) and the fraction of time steps that we can simulate compared to the real case (10^7 simulated

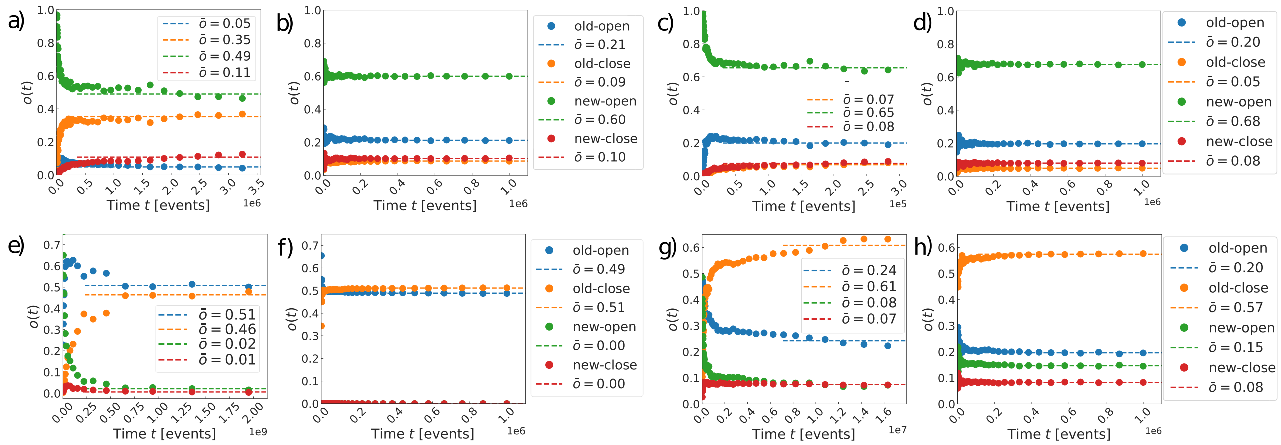

Fig. 6 Dynamics of links formation: empirical data and model predictions. The plot of the empirical data and model results for the fraction of activated links that are: old (i.e., they have been active before) and open (they close no triangle) $OO(t)$ (blue markers), old and close (i.e., they close at least one triangle) $OC(t)$ (orange markers), new (i.e., they get activated for the first time) and opens $NO(t)$ (green markers), and new and close $NC(t)$ (red markers). In each case, we show the temporal behaviour (markers), and the asymptotic value (dashed lines, the asymptotic value is reported in the legends). Panels refer to (**a**) American physical society (APS) and (**b**) its model, (**c**) 1-link subsampling of APS and (**d**) the synthetic model, (**e**) the mobile phone network (MPN) dataset and (f) it is corresponding numerical one, and (**g**) the Twitter mentions network TMN dataset with (**h**) the synthetic one.

steps versus the $2 \cdot 10^9$ real events). Nevertheless, the overall shape of the communities size distribution is also reproduced in this case.

We then check if the generated networks also reproduce the empirical core-periphery structure. Intuitively, a core-periphery structure corresponds to a network with communities composed by a densely connected set of nodes (the core) to which other nodes sparsely connected amongst them are linked (the periphery, see Methods for details). We report the results in Fig. 5f where we show the distribution of the fraction of core nodes detected for each community. As one can see, the model correctly reproduces even the proportion of core and periphery nodes in communities.

Finally, passing to the microscopic dynamics, we report the temporal evolution of the number of events allocated by nodes toward new or old links insisting or not on triangles (the categories *OO*, *OC*, *NO*, and *NC* defined above and in the “Methods”). Again, we note that, while the original dataset of APS shows a behaviour not well reproduced by the model, the latter very nicely predicts the behaviour featured by the 1-link sub-sampled APS dataset, as well as the behaviour featured by the other two datasets (Fig. 6).

These additional results overcome some of the limitations of previous models and approaches. For example, in[34] the community structure has to be fixed from the beginning of the simulation, whereas other models where the community structure emerges from the network growth do not account for either the edges weights[9,13,33], the network dynamics[36] or the topological correlations of weights and overlap[35]. Moreover, none of the network dynamic models we are aware of can reproduce this large set of dynamical and topological observables at once. Either they focus on the link weights allocation[28], on the activation patterns of nodes[35] or they do not account for global observables such as community structures and modularity at all[36].

Optimal exploration strategies. Let us finally note that the optimal parameters values found for each dataset draws some meaningful insights on the microscopic mechanisms driving the exploration of the social space in the different context of each social network. In the TMN case, we find $R = 1$, so that the reinforcement and the novelty exchange processes equally influence the single agents’ exploration process: this is reasonable in a system where new connections require little effort from the user. Moreover, the strategy $s =$ WSW (weighted sampling with withdrawal) with $\nu = 5$ is the one that better describes the empirical data: users select new accounts to mention by sampling from the past interactions of the alters they are connecting with proportionally to the number of their past interactions.

On the other hand, in the MPN case, the best fit is obtained for $R = 3$ and $\rho = 21$. The system dynamic is dominated by reinforcement processes that tend to reinforce links that are first established and inhibiting the creation of new edges. In this case, the best fit with the $s =$ ASW memory buffer sampling strategy highlights that individuals share their last $\nu + 1 = 8$ contacts, thus spreading copies of recently contacted IDs rather than the most contacted ones. Notice that the last contacted $\nu + 1$ IDs may, in general, be different from the most representative IDs within the urn. The asymmetric nature of the ASW strategy indicates that users actively exploring new connections update their memory buffer, whereas nodes passively participating in communication tend to conserve their previous memory buffers.

Finally, in the APS case, we find an extremely exploratory dynamics characterised by a relatively low $R = 0.4$, i.e., a relatively high ν. This finding is symptomatic of a dynamics where the exploration of the social space overtakes the reinforcement of existing connections. A possible explanation lies in a large number of students and researchers authoring a few papers before quitting academia, providing a constant influx of new potential connections to be explored by senior researchers. The *SSW* optimal sampling strategy reveals that authors tend to share their last $\nu + 1 \simeq 16$ people they have been collaborating with, implying a preference to recommend recently active connections to new collaborators. Moreover, this strategy also catches the intrinsic symmetric nature of the co-authorship interaction, as both co-authors update their buffers of potential new collaborators. Further confirmation of this is that the sub-sample of the APS dataset is best fitted by a comparable $R = 0.33$ but with an ASW strategy $s =$ ASW, where only the node actively engaging the interaction updates its memory buffer.

Conclusion

In this work, we proposed a theoretical model of social exploration to explain the birth and evolution of social networks. The theory is based on the adjacent possible framework and builds on a recently introduced mathematical formalisation of its conditional expansion. In this framework, the creation of new social bonds is the outcome of an exploration process unfolding on the space of possible new acquaintances, whose boundaries change while people explore them.

Without relying on unnecessary assumptions, our theoretical model builds on the adjacent possible space expansion and microscopic evolution rules that let emerge both microscopic and macroscopic features of real-world social networks. We compared the predictions with the empirical data from three diverse social networks: the network of mentions between Twitter users, the network of co-authorship of the APS, and a mobile phone-call network. The agreement between theory and data is surprisingly good. On the macroscopic side, the model reproduces the main static and dynamic features of those social networks: the broad distribution of degree and activities, the average clustering coefficient, and the innovation rate at the global and local levels. At the microscopic level, the most striking feature captured is the probability for an individual, with already k connections in its local network, to acquire a new acquaintance. The model also captures the topological correlations, the modular structure and the core-periphery organisation of nodes. Besides, the model is able to grasp the temporal-evolution of real-world systems at very different scales, from the local exploit/explore mechanisms of single agents to the global organisation of the network in communities of coherent users. To the best of our knowledge, this is the first model that reproduces these features without superimposing a heterogeneous fitness distribution to nodes, a specific rule to choose whether to interact with a new or old alter or to set the modularity of the initial network manually. Moreover, it is the first attempt to reproduce both the network growth (in terms of nodes and edges) and its dynamics (the sequence of activation events of the edges in time) at once, giving results in excellent agreement with the empirical case.

Besides being able to capture very complex features of social networks quantitatively, our theory also allows us to deepen our understanding of the microscopic mechanisms shaping the propensity of people to reinforce old contacts or establish new ones. For instance, in the Twitter mentions network, we find the exploration and reinforcement processes to be of equal importance. Moreover, when getting in contact with new alters, users share a sample of their most common contacts as new potential connections. On the other end of the spectrum, in the mobile phone-calls network, people reinforce their existing bonds more than they explore new ones. When suggesting new potential contacts to others, people tend to exchange their most recent contacts, rather than their most common ones. Finally, the network of scientific co-authorship of the APS journals features the most exploratory dynamics, with new connections massively expanding the adjacent possible space of a single node. In this case, people preferably share their last contacts, and the optimal synthetic update procedure is symmetrical, correctly reproducing the intrinsic symmetric nature of the interactions.

The theoretical framework proposed here is, of course, open to possible improvements. First, the simulated dynamics describes the evolution of a system from its outset. The initial conditions set here could be far from those of the real-world systems considered. Despite the excellent agreement with empirical data, a more comprehensive study on the dependence of the system evolution on the initial state is in order. Other generalisations could concern the possibility to remove links[52] or to generalise the model to have a separate rate of nodes and links entrance in the system. Finally, our modelling scheme does not account for effects connected to semantics or affinity between people. For instance, it seems reasonable to assume that people create bonds and interact based on shared interests or their level of homophily[53,54]. The generality of the approach presented here will make the extension of the theoretical framework desirable and possible along these lines. Also, we restricted to the case when nodes exchange their memory buffer only on their first encounter, whereas other strategies may be considered. Another important extension is to test the predictive potential of the model fitting it on a subset of the points and test its forecasting ability to determine which nodes will be in contact (and with which strength) in the future.

We believe that the presented framework, together with its predictions validated on real-world social networks, represents a valuable step toward understanding the processes underlying the birth and evolution of social networks. It further creates an important bridge between network theory and urns models, opening the way to constructive contamination between the two fields and full exploitation of results derived for stochastic processes relevant in innovation dynamics[50,51,55]. This development, in turn, unlocks the possibility to grasp the very essence of social interactions and allows for the design of efficient and informed policies to address crucial challenges dealing with collective processes ongoing in social networks, such as the spread of diseases and online misinformation.

Methods

Data and code. Here, we summarise the empirical data used to test the predictions of our theory. These are three different real-world social networks: (i) the APS co-authorship network generated by all the papers published in all the APS journals from January 1970 to December 2006. (ii) TMN logging all the mentions between users recorded between January and September 2008. (iii) MPN recording the calls between users of a national provider in an undisclosed European country between January and July 2008. We refer also to the Supplementary Note 2 for details. These datasets represent diverse contexts of social interactions, making them an ideal set of empirical observations to test the universality of our model. In particular, the APS dataset describes the undirected interactions of co-authors of scientific papers[56–59]. Here interactions have a high cost in terms of time and resources. The TMN dataset reports the directed citations of a user i citing a user j (that corresponds to an edge from i to j) between users of the micro-blogging platform, in which interactions are requiring few resources and can be virtually established from and to any node in the network[60]. Finally, the MPN dataset lies somewhere in between: communication is not as cheap as in the TMN but still easier than in the APS case[8]. Also, the network may be not single scoped for the users taking part in it: some of them may use it to call close contacts whereas others may use the phone for business reasons[17,25]. Let us also note that the TMN and APS datasets account for the growth of the two systems since their onset. Indeed, the effective onset of user adoption for Twitter occurred during 2008[61], whereas the APS created the majority of its journals in 1970. This circumstance ensures a unique testbed for a model of network growth. On the other hand, the MPN situation is more subtle as we have only a limited observation window on a system that underwent a long evolution period beforehand. Summarising, the three datasets used in the study are:

- The co-authorship networks found in the Journals of the APS[59] covering the period between January 1970 and December 2006 and containing 301,236 papers written by 184,583 authors that are connected by 995,904 edges.
- TMN, containing all the mention events exchanged by users from January to September 2008. The network has 536,210 nodes performing about 160 M events and connected by 2.6 M edges;
- MPN composed of 6,779,063 users of a single operator with about 20% market share in an undisclosed European country from January to July 2008. The datasets contain all the phone calls to and from company users, thus including the calls towards or from 33,160,589 users in the country connected by 92,784,825 edges.
- The synthetic simulations have been run for $T = 10^6$ evolution steps for configurations with $R \leq 1$, $T = 5 \cdot 10^7$ otherwise. We performed ten independent simulations for each set of parameters. We define as the best set of parameters the set featuring the lowest average cost function on those ten replicas. In the figures, we report the figures and behaviour of the run that gave the minimum cost function among these ten runs.

The code used to run the simulations, all the analysis code, as well as the synthetic data analysed, are available in[62]. Due to data policies and IPR, we cannot share the MPN data, while the APS data are from the work in[59] and the TMN data are made available on figshare (see link in the Data availability section) . Finally, let us note that we used these three datasets despite the wealth of longitudinal network datasets present in the literature because we needed the complete sequence of events to measure and fit all the dynamical observables.

Asymptotic behaviour of the system. In this work, we leverage on a previous analysis performed on the same datasets as found in[25]. Specifically, we measure the strengthening probability $p_i(k)$, i.e., the probability for an individual i who already contacted k distinct individuals in the past to contact a new one (i.e., a new node of the network). To average this probability on homogeneous classes of people, we divide the nodes in $e = 1, \ldots, E$ classes depending on their time of entrance in the system, t_e, and their final degree k_e, one class for each combination of t_e and k_e. The functional form of the probability $p_e(k)$ is found to depend on class e of the nodes as $p_e(k) = (1 + k/c_e)^{-\beta}$ with a single overall β exponent and a distributed reinforcement constant c_e. The latter is fixed by computing, for each k the ratio between the number of events resulting in a degree increase from k to $k + 1$ performed by the nodes belonging to the class e and the total number of events performed by the same set of nodes at degree k (thus including the ones toward already contacted nodes). As for the growth of the average degree $\langle k(t_e, t)\rangle$, we measure the average degree at time $t > t_e$ for all the nodes belonging to the class with entrance time t_e. In this way, we are defining a new set of classes only defined in terms of the entrance time t_e. Asymptotic behaviour is found to be $\langle k(t_e, t)\rangle \propto t^q$.

Model cost function. We ran the model at different values of R and ρ for each one of the six sample strategies s (see Supplementary Note 3 for details). For each dataset d we select the configuration that best fit the data by minimising the cost $S_d(\rho, R, s)$ that reads

$$S^d(\rho, R, s) = \sum_{i=1}^{8} \frac{|o_i^d - \tilde{o}_i(\rho, R, s)|}{\sigma_i^d},$$

where o_i^d and σ_i^d are the value and uncertainty on the ith observable of the empirical dataset and $\tilde{o}_i(\rho, R, s)$ is the value of the same observable measured in the simulations with configuration (ρ, R, s). The eight selected observables are: (1) the exponent γ leading the growth of the number of edges $E(t) \propto t^\gamma$, (2) the optimal β measured in the strengthening function $p(k)$, (3) the average clustering coefficient c, (4) the exponent leading the growth of the average degree per node class $\langle k(e, t)\rangle \propto t^q$, (4)–(8) the fractions OO, OC, NO, NC of events allocated toward old/new link insisting or not on an open/closed triangle.

APS subsampling. In the APS dataset, we transform each paper published by n authors in a sequence of $E = n(n - 1)$ events with all the possible links between all the ordered couples of co-authors. We then sample l links over the E possible links for each paper to be inserted in the total sequence $\mathcal{S}$. The results reported in the main text refer to $l = 1$, and the reader can refer to the Supplementary Note 4 for results with different values of l and different strategies of subsampling (number of sampled links proportional to E).

Modularity and core-periphery measures. Communities and modularity values have been found using the Infomap algorithms in the Python `iGraph` module[63], using the edges weights w_{ij} as the weight parameters and with ten trials. For computational reasons, we restricted ourselves to a sub-graph induced by a subsample of 100,000 nodes in all the analysed cases (both empirical and synthetic). These sub-samples have been determined using the sampling method found in[64] in its deterministic version. Indeed, this method returns a set of nodes whose subnetwork best reproduces the modularity, clustering and overlap features of nodes in a network. In the core-periphery analysis, the algorithms find another set of communities as this algorithm uses a configuration model as its null model. However, we show in the Supplementary Note 5 that the $P(r)$ distribution of community sizes is consistent with the one found by the Infomap algorithm. For the core-periphery analysis, we run the model introduced in[65] with its C++ implementation found in[66] against the same sub-graphs of the previous analysis. Also in this case we use the edges' weights w_{ij} as the weight passed to the algorithm.

Events on new-old and open-closed edges. We count, for each logarithmically spaced time interval, the number of events happening on edges that are either old (already activated in the past), new (being activated now) and that happen to close a triangle (closed) or not (open). These four categories are then: the OO, OC, NO, and NC that we define as the fraction of events falling in each category per time range in the asymptotic limit of the system evolution—i.e., after 60% of the events passed.

Data availability

Aggregated data used to produce all the population-level figures in the paper have been deposited on figshare at https://doi.org/10.6084/m9.figshare.13308428. This repository also contains the raw data of Twitter and APS datasets. The mobile phone network dataset was not directly accessible to the authors as it was analysed by uploading the analysis scripts to Marton Karsai at ENS Lyon, France, which we thank for the help provided.

Code availability

All the code used to analyse and produce the paper figures is available on GitHub: the code used to analyse the data is in https://github.com/ubi15/pytvn while the code to run the simulations and analyse the output is on https://github.com/ubi15/pyUrns.

Received: 21 July 2020; Accepted: 6 January 2021;
Published online: 16 February 2021

References

1. Marsili, M., Vega-Redondo, F. & Slanina, F. The rise and fall of a networked society: a formal model. *Proc. Natl Acad. Sci. USA* **101**, 1439–1442 (2004).
2. Granovetter, M. *Getting a Job: A Study of Contacts and Careers. Sociology* (University of Chicago Press, 1995) https://books.google.fr/books?id=R7-w4BLg7dAC.
3. Sekara, V., Stopczynski, A. & Lehmann, S. Fundamental structures of dynamic social networks. *Proc. Natl Acad. Sci. USA* **113**, 9977–9982 (2016).
4. Kossinets, G. & Watts, D. J. Empirical analysis of an evolving social network. *Science* **311**, 88–90 (2006).
5. Perra, N., Gonçalves, B., Pastor-Satorras, R. & Vespignani, A. Activity driven modeling of time varying networks. *Sci. Rep.* **2**, 469 (2012).
6. Davidsen, J., Ebel, H. & Bornholdt, S. Emergence of a small world from local interactions: Modeling acquaintance networks. *Phys. Rev. Lett.* **88**, 128701 (2002).
7. Jin, E. M., Girvan, M. & Newman, M. E. J. Structure of growing social networks. *Phys. Rev. E* **64**, 046132 (2001).
8. Onnela, J.-P. et al. Structure and tie strengths in mobile communication networks. *Proc. Natl Acad. Sci. USA* **104**, 7332–7336 (2007).
9. Bianconi, G., Darst, R. K., Iacovacci, J. & Fortunato, S. Triadic closure as a basic generating mechanism of communities in complex networks. *Phys. Rev. E* **90**, 042806 (2014).
10. Kumpula, J. M., Onnela, J.-P., Saramäki, J., Kaski, K. & Kertész, J. Emergence of communities in weighted networks. *Phys. Rev. Lett.* **99**, 228701 (2007).
11. Kumpula, J. M., Onnela, J.-P., Saramäki, J., Kertész, J. & Kaski, K. Model of community emergence in weighted social networks. *Comput. Phys. Commun.* **180**, 517–522 (2009).
12. Granovetter, M. S. The strength of weak ties. *American journal of sociology* **78**, 1360–1380 (1973).
13. Lambiotte, R., Krapivsky, P. L., Bhat, U. & Redner, S. Structural transitions in densifying networks. *Phys. Rev. Lett.* **117**, 218301 (2016).
14. Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep.* **519**, 97–125 (2012).
15. Moinet, A., Starnini, M. & Pastor-Satorras, R. Burstiness and aging in social temporal networks. *Phys. Rev. Lett.* **114**, 108701 (2015).
16. Saramäki, J. et al. Persistence of social signatures in human communication. *Proc. Natl Acad. Sci. USA* **111**, 942–947 (2014).
17. Miritello, G., Lara, R., Cebrian, M. & Moro, E. Limited communication capacity unveils strategies for human interaction. *Sci. Rep.* **3**, 1950 (2013).
18. Miritello, G., Moro, E. & Lara, R. Dynamical strength of social ties in information spreading. *Phys. Rev. E* **83**, 045102 (2011).
19. Vestergaard, C. L., Génois, M. & Barrat, A. How memory generates heterogeneous dynamics in temporal networks. *Phys. Rev. E* **90**, 042805 (2014).
20. Barrat A., Barthlemy M. & Vespignani A. *Dynamical Processes on Complex Network*. 1st edn (Cambridge University Press: New York, NY, USA, 2008).
21. Karsai, M., Perra, N. & Vespignani, A. Time varying networks and the weakness of strong ties. *Sci. Rep.* **4**, 4001 (2014).
22. Karsai, M. et al. Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E* **83**, 025102 (2011).
23. Ubaldi, E., Vezzani, A., Karsai, M., Perra, N. & Burioni, R. Burstiness and tie activation strategies in time-varying social networks. *Sci. Rep.* **7**, 46225 (2017).
24. Barrat, A., Fernandez, B., Lin, K. K. & Young, L.-S. Modeling temporal networks using random itineraries. *Phys. Rev. Lett.* **110**, 158702 (2013).
25. Ubaldi, E. et al. Asymptotic theory of time-varying social networks with heterogeneous activity and tie allocation. *Sci. Rep.* **6**, 35724 (2016).
26. Topirceanu, A., Udrescu, M. & Marculescu, R. Weighted betweenness preferential attachment: a new mechanism explaining social network formation and evolution. *Sci. Rep.* **8**, 1–14 (2018).
27. Zuev, K., Boguná, M., Bianconi, G. & Krioukov, D. Emergence of soft communities from geometric preferential attachment. *Sci. Rep.* **5**, 9421 (2015).
28. Laurent, G., Saramäki, J. & Karsai, M. From calls to communities: a model for time-varying social networks. *Eur. Phys. J. B* **88**, 1–10 (2015).
29. Cattuto, C., Barrat, A., Baldassarri, A., Schehr, G. & Loreto, V. Collective dynamics of social annotation. *Proc. Natl Acad. Sci. USA* **106**, 10511–10515 (2009).
30. Colman, E. & Rodgers, G. Local rewiring rules for evolving complex networks. *Physica A* **416**, 80–89 (2014).

31. Holme, P. & Ghoshal, G. Dynamics of networking agents competing for high centrality and low degree. *Phys. Rev. Lett.* **96**, 098701 (2006).
32. Rosvall, M. & Sneppen, K. Modeling dynamics of information networks. *Phys. Rev. Lett.* **91**, 178701 (2003).
33. Hébert-Dufresne, L., Laurence, E., Allard, A., Young, J.-G. & Dubé, L. J. Complex networks as an emerging property of hierarchical preferential attachment. *Phys. Rev. E* **92**, 062809 (2015).
34. Karan, R. & Biswal, B. A model for evolution of overlapping community networks. *Physica A* **474**, 380–390 (2017).
35. Kasper, P. et al. Modeling user dynamics in collaboration websites (eds Ghanbarnejad, F., Saha Roy, R., Karimi, F., Delvenne, J.-C. & Mitra, B.) *Dynamics on and of Complex Networks III*, 113–133 (Springer International Publishing, Cham, 2019).
36. Overgoor, J., Benson, A. & Ugander, J. Choosing to grow a graph: modeling network formation as discrete choice. In *Proc. World Wide Web Conference*, WWW'19, 1409–1420 (Association for Computing Machinery, New York, NY, USA, 2019). https://doi.org/10.1145/3308558.3313662.
37. Kauffman, S. *The Origins of Order: Self-organization and Selection in Evolution* (Oxford University Press, 1993). https://books.google.fr/books?id=lZcSpRJz0dgC.
38. Kauffman, S. & Santa, N. Fe Institute. Santa Fe, *Investigations: The Nature of Autonomous Agents and the Worlds they Mutually Create*. SFI working papers (Santa Fe Institute, 1996). https://books.google.fr/books?id=IgiOPwAACAAJ.
39. Kauffman, S. A. *Investigatios* (Oxford University Press, 2000).
40. Tria, F., Loreto, V., Servedio, V. D. P. & Strogatz, S. H. The dynamics of correlated novelties. *Sci. Rep.* **4**, 5890 (2014).
41. Loreto, V., Servedio, V. D. P., Strogatz, S. H. & Tria, F. *Dynamics on Expanding Spaces: Modeling the Emergence of Novelties*, 59–83 (Springer International Publishing, Cham, 2016). https://doi.org/10.1007/978-3-319-24403-7_5.
42. Pólya, G. Sur quelques points de la théorie des probabilités. *Ann. Inst. H. Poincaré* **1**, 117–161 (1930).
43. Mahmoud, H. *Pólya urn models* (Chapman and Hall/CRC, 2008).
44. Monechi, B., Ruiz-Serrano, A., Tria, F. & Loreto, V. Waves of novelties in the expansion into the adjacent possible. *PLoS ONE* **12**, 1–18 (2017).
45. Saracco, F., Di Clemente, R., Gabrielli, A. & Pietronero, L. From innovation to diversification: a simple competitive model. *PloS ONE* **10**, e0140420 (2015).
46. Jo, H.-H., Karsai, M., Kertész, J. & Kaski, K. Circadian pattern and burstiness in mobile phone communication. *New J. Phys.* **14**, 013055 (2012).
47. Taylor, L. Aggregation, variance and the mean. *Nature* **189**, 732 (1961).
48. Eisler, Z., Bartos, I. & Kertész, J. Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv. Phys.* **57**, 89–142 (2008).
49. Gerlach, M. & Altmann, E. G. Scaling laws and fluctuations in the statistics of word frequencies. *New J. Phys.* **16**, 113010 (2014).
50. Tria, F., Crimaldi, I., Aletti, G. & Servedio, V. D. P. Taylor's law in innovation processes. *Entropy*. **22** (2020). https://www.mdpi.com/1099-4300/22/5/573.
51. Tria, F., Loreto, V. & Servedio, V.D.P. Zipf's, heaps' and Taylor's laws are determined by the expansion into the adjacent possible. *Entropy*. **20** (2018). https://www.mdpi.com/1099-4300/20/10/752.
52. Murase, Y., Jo, H.-H., Török, J., Kertész, J. & Kaski, K. Modeling the role of relationship fading and breakup in social network formation. *PLoS ONE* **10**, 1–14 (2015).
53. Murase, Y., Jo, H.-H., Török, J., Kertész, J. & Kaski, K. Structural transition in social networks: the role of homophily. *Sci. Rep.* **9**, 1–8 (2019).
54. Asikainen, A., Iñiguez, G., Ureña-Carrión, J., Kaski, K. & Kivelä, M. Cumulative effects of triadic closure and homophily in social networks. *Sci. Adv.* **6**, eaax7310 (2020).
55. Pitman, J. *Combinatorial Stochastic Processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII* (Springer, 2002).
56. Newman, M. E. The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA* **98**, 404–409 (2001).
57. Newman, M. E. Scientific collaboration networks. i. Network construction and fundamental results. *Phys. Rev. E* **64**, 016131 (2001).
58. Newman, M. E. Scientific collaboration networks. ii. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 016132 (2001).
59. Radicchi, F., Fortunato, S., Markines, B. & Vespignani, A. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E* **80**, 056103 (2009).
60. Gonçalves, B., Perra, N. & Vespignani, A. Modeling users' activity on twitter networks: validation of dunbar's number. *PloS ONE* **6**, e22656 (2011).
61. Bufferapp, how twitter evolved from 2006 to 2011. https://blog.bufferapp.com/how-twitter-evolved-from-2006-to-2011 (2016) accessed 01 December 2018.
62. Ubaldi, E. Pyurns. https://github.com/ubi15/pyUrns (2019).
63. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Int. J.* Complex Syst. 1695 (2006). http://igraph.org.
64. Maiya, A. S. & Berger-Wolf, T. Y. Sampling community structure. In *Proc. 19th International Conference on World Wide Web*, WWW'10, 701–710 (Association for Computing Machinery, New York, NY, USA, 2010). https://doi.org/10.1145/1772690.1772762.
65. Kojaku, S. & Masuda, N. Core-periphery structure requires something else in the network. *N. J. Phys.* **20**, 043012 (2018).
66. Kojaku, S. & Masuda, N. Pyurns. https://github.com/skojaku/km_config (2020).

Acknowledgements

The authors wish to thank M. Karsai for useful comments and for granting us access to the mobile phone dataset and V.D.P. Servedio for inspiring discussions and relevant suggestions.

Author contributions

E.U., F.T., V.L., and R.B. conceived the research. E.U. developed the code, performed the simulations and analysed the data; E.U., F.T., V.L., and R.B. wrote and reviewed the paper and gave final approval for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-021-00527-1.

Correspondence and requests for materials should be addressed to F.T.

Reprints and permission information is available at http://www.nature.com/reprints