Understanding mobility in a social petri dish

Supplementary Information

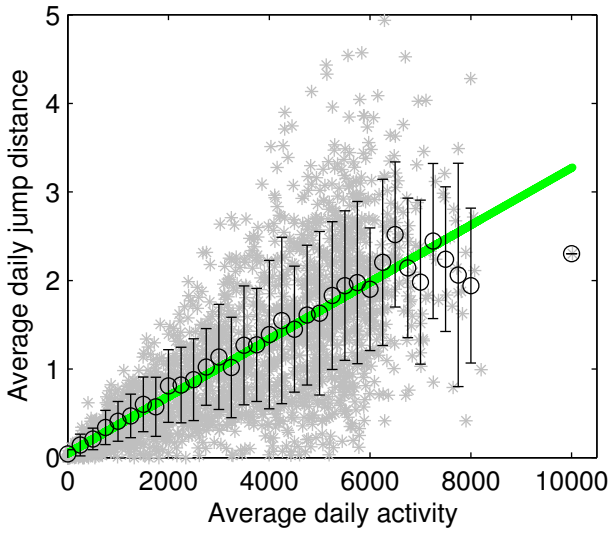
Michael Szell, Roberta Sinatra, Giovanni Petri, Stefan Thurner, Vito Latora

CONTENTS

S1. ACTIVITY OF INHABITANTS IN THE ONLINE SOCIETY OF PARDUS

Pardus (www.pardus.at) is a browser-based Massive Multiplayer Online Game (MMOG) in a science-fiction setting, online since September 2004. Every player owns an account with up to one *character* per game universe. A character is a pilot owning a spacecraft with a certain cargo capacity, roaming the virtual universe trading commodities, socializing, and much more, 'to gain wealth and fame in space' (http://www.pardus.at/index.php?section=about). The main component of Pardus consists of trade with a society of players heavily driven by social factors such as friendship, cooperation or competition [1, 2]. There is no explicit 'winning' in Pardus as there is no inherent set of goals. Pardus is a *virtual world* whose gameplay is based on socializing and role-playing, with interaction between players and non-player characters as its core elements [3].

Space in Pardus is two-dimensional. There are three topologically identical but independent game universes. Between universes it is impossible to move, trade, or exchange game money. A universe is divided into 400 sectors, each sector consisting of 15×15 fields on average. Fields are the smallest units of space and are displayed as square images in-game. They form a square grid on which continuous movement is possible by clicking on the desired destination field within the space chart. This chart is a 7×7 fields cut-out of the universe visible to every player centred on their current position. A sector's boundary is impenetrable; moving between nearby sectors is possible only by tunneling through field objects called *wormholes*. Moving between sectors which are not close-by on the universe map is possible via field objects called *X-holes* and *Y-holes*, which are equivalent to wormholes but require a higher cost of action points to traverse (see below). Sectors are organized in *clusters*, composed on average of 20 neighbouring sectors. The typical spatial range of activities of a character is usually confined to one cluster for several weeks or longer. For more details on player interaction and means of movement in Pardus see [1, 2].



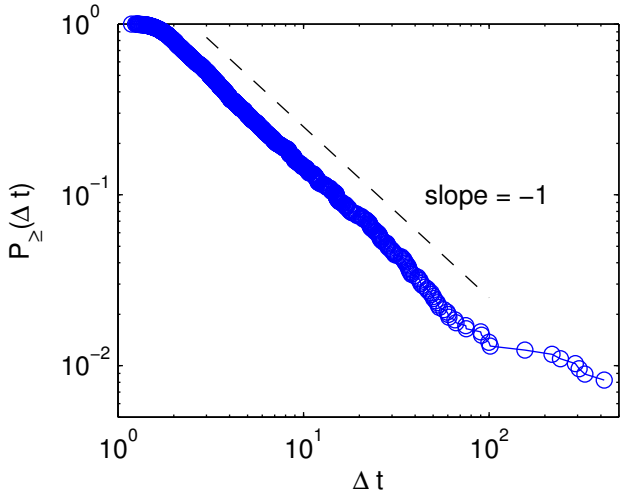
Supplementary Figure 1. Average daily jump distance versus average daily activity as measured in Action Points (APs) spent, for each player. Circles and error bars are binned averages with standard deviations. The green line is a least-squares fit, making a trend visible between mobility and activity. High activity is necessary, but not sufficient, for high mobility: in order to have a high mobility, a player needs to spend many APs (there are no data points in the top left part), but there exists a number of players with low mobility who spend a high number of APs (there are data points in the lower right part).

Every game action carried out by a player (trade, movement, etc.), except for communication activities (using the game's chat channels, writing messages, posting in the game's forums), costs a certain amount of *action points* (APs). These points can not exceed a maximum value of 5000 or slightly more (different characters can have slightly different maximum AP values, but in a neglectable order). For characters owning less APs than their maximum, every 6 minutes 24 APs are automatically regenerated, i.e. 5760 APs per day. Once a player's character is out of APs, she has to wait to be able to play on. As a result the typical Pardus player logs in once a day to spend all her APs on several activities within a few minutes (for each character/universe). This makes APs, the game's unit of time, the most valuable factor: those players who use their APs most efficiently can experience the fastest progress or earn the highest profits. Note that it is up to the player to spend (or not spend) her daily APs gained, and distribute them freely on any of the possible types of AP-consuming activities. In a rough estimation, it takes around $1,500$ APs to

traverse a typical cluster, and approximately 10, 000 (around 2 days worth of APs) to traverse the whole universe.

Figure 1 shows the average daily jump distance versus the average daily number of APs spent for each player. The number of APs used per day is distributed heterogeneously, i.e. some players are much more active than others. In general, the more APs a player uses, the more she travels. However, high activity is necessary but not sufficient for high mobility: in order to travel far, a player needs to spend many APs (corresponding to a lack of data points in the top left part of the figure), but there exists a number of players with low mobility who spend a high number of APs (there exist data points in the lower right part of the figure).

Another quantity which indicates the extent of activity in mobility is the average waiting time $\overline{\Delta\, t}$ of a single player, i.e. how many days a single player stays on a sector on average. The cumulative distribution of these average waiting times of all players, Supplementary Fig. 2, displays a power law with slope -1, showing further the heterogeneity of player activity in respect to mobility, motivating the transformation from real-time t to jump-time τ in the main text for determining the distribution of first return times $P_{\leftarrow}(\tau)$.

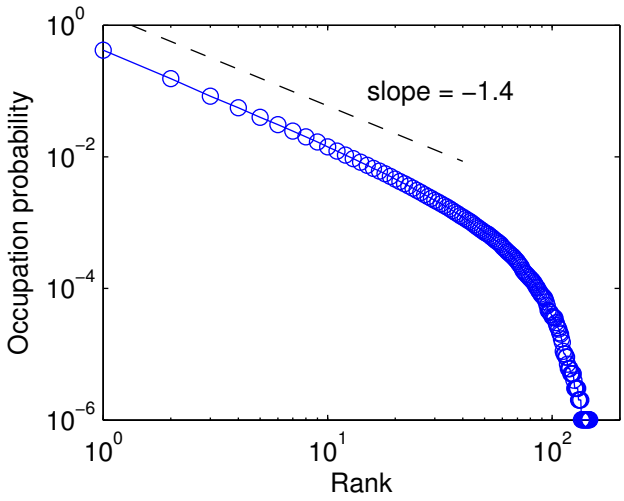


Supplementary Figure 2. Cumulative distribution of average waiting times $\overline{\Delta\, t}$ of single players. The dashed line is a power law with slope -1, which fits well the distribution. The fat-tailed nature of this distribution shows the heterogeneity in movement behaviour of players.

Supplementary Fig. 3 further shows the distribution of occupation probabilities, averaged over the distributions of all single players, ordered by rank. Single players may have different locations which they prefer, however the averaged distribution over these single, rank-ordered occupation probabilities shows the common pattern of significantly prefering the stay in certain locations over other locations.

S2. DATASET DESCRIPTION

We focus on one of the three *Pardus* universes, *Artemis*, because complete information from the beginning of this universe is available, and there exists a large number of players due to Artemis being a free-to-play universe [1, 2]. To make sure we only consider active players, we select all players who exist in the game between the days 200 and 1200. We discard the first 200 days because social networks between players of *Pardus* have shown aging effects in the beginning of the universe, i.e. there seems to exists a transient phase in the development of the society, possibly affecting mobility, which we would like to avoid considering [1]. This cut selects 1458 players active over a time-period of 1000 days. The field-ID (position within a sector) of these players is logged every day at 05:35 GMT. From these field-IDs, we select the corresponding sector-IDs, i.e. the player positions on the nodes of the universe network, leaving us with a 1458×1000 mobility matrix containing 1458 sequences each consisting of 1000 sector-IDs. Note that we use daily snapshots of position data since we are interested in the long-time mobility and not in the detailed paths the players take during the typically few minutes of their daily navigations. The effects of this 'systematic coarse-graining' seem tame compared to biases occurring for example in mobile phone data, where location information is usually only available at the heterogeneously distributed times when a mobile phone is used, and locations, hence distances, have to be triangulated via heterogeneously scattered mobile phone towers [4]. These issues inherent in previously studied



Supplementary Figure 3. Distribution of occupation probabilities, averaged over the distributions of all single players, ordered by rank. The dashed line is a power law with slope −1.4, which fits well the distribution. The fat-tailed nature of this distribution shows how players have clear preferences to stay much longer on or to visit much more frequently certain locations than others.

mobile phone data may become resolved in future studies, considering recent disclosures that exact location data (GPS) is constantly recorded by a number of mobile phone devices [5].

S3. UNIVERSE PROPERTIES

Topologically, the universe network is sparse, displaying an average degree of $\bar{k} = 2.9$ with $N = 400$ nodes and $K = 1160$ links. It is highly clustered (clustering coefficient $C = 0.089$, which is 12 times larger than C_r, the average clustering coefficient for corresponding random networks, i.e. networks having the same number of nodes and the same average degree as the universe network), but is not small-world [6]: the characteristic path length $\bar{L} = 11.89$ is more than twice L_r, the average path length on a random graph with the same number of nodes [7] and the same average connectivity, and relative to the number of nodes the diameter $d_{\max} = 27$ is large. Excluding X-hole and Y-hole links, the universe is a planar network, i.e. it has no intersecting links. In this sense it is similar to Euclidean networks such as street networks, with the difference that the length of links has no effect in Pardus. Locally, the universe is lattice-like, which is reflected in the relatively high local efficiency $E_{\text{loc}} = 0.80$ [8] and the narrow degree distribution (almost all sectors have degree 2, 3, or 4).

Besides the clusters, which are defined externally, there is the possibility for players to build obstacles called 'military outposts' blocking almost any region of space from being entered by single or groups of other players. This option adds player-driven 'political borders' to the natural ones, just as borders of nations often constitute significant limitations to mobility of human beings. Such player-built obstacles can be taken down by other players through the use of (usually organized) hostile force. Considering the political dynamics of military outposts is beyond the scope of this article. Note, however, that often groups of players construct military outposts at the borders of clusters, possibly further adding to the mobility-hindering effect of clusters quantified in the main text.

S4. COMMUNITY DETECTION

Since the Pardus universe is divided in a number of *clusters*, defining political, independent units in the game universe, it is reasonable to expect (and is actually shown in the main text) that the mobility patterns of players are influenced by such borders. At the same time, the topology of the Pardus universe itself might affect the mobility patterns. In order to investigate the importance of these two elements, one needs to compare the topological modules that can be extracted from the adjacency matrix A or distance matrix D, with the dynamical communities emerging from the collective movement behaviour of players.

N	400
K	1160
$\bar{k}$	2.9
C	0.089
$C/C_{\rm r}$	12.33
L	11.89
$L/L_{\rm r}$	2.11
$E_{\rm loc}$	0.80
$E_{\rm glob}$	0.03
$d_{\rm max}$	27

Supplementary Table I. Network properties of the Pardus universe: number of nodes N, number of (undirected) links K, average degree $\bar{k}$, clustering coefficient C, clustering coefficient to corresponding coefficient of random graph $C/C_{\rm r}$, average geodesic L, average geodesic to corresponding average geodesic of random graph $L/L_{\rm r}$, local efficiency $E_{\rm loc}$, global efficiency $E_{\rm glob}$, diameter $d_{\rm max}$.

.

At the sector level, the Pardus universe is a directed weighted network with $L = 1160$ links and $N = 400$ nodes. The majority of links are *wormholes* ($\sim 95\%$), mutual links that connect nearby nodes (see Fig. 1 in the main text) and have a small traveling cost (in terms of APs). The long-range links in Fig. 1 in the main text instead represent *X-holes* and *Y-holes*. Players moving along such links incur significantly higher traveling costs than in the case of wormholes. Since X/Y-holes may be only scarcely used in-game, in addition to studying the complete directed weighted adjacency matrix, A, we also study the adjacency matrix $A^{\rm wh}$ where X/Y-holes were removed, yielding a symmetric and unweighted network. Finally, we consider the weighted network $D^{\rm inv}$, defined element-wise as $d_{ij} = d(i,j)^{-1}\ \forall\ i \neq j$ and $d_{ii} = 0\ \forall i$, where $d(i,j)$ is the shortest path distance on the Pardus network.

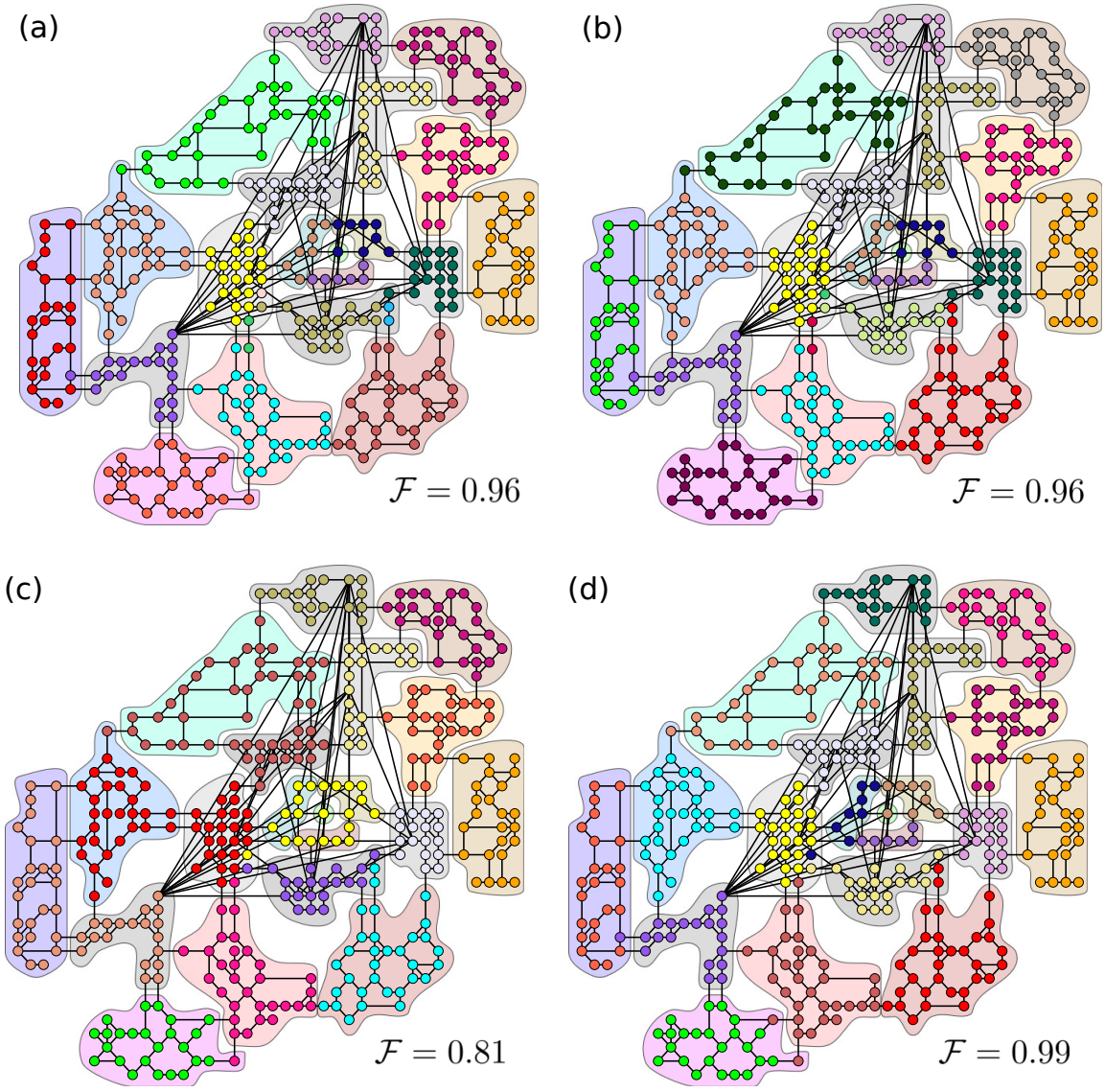
The player dynamics was studied at the aggregate level through the transition count matrix M and the normalized transition matrix Π. Each element m_{ij} of M corresponds to the total number of times a player was found at position i at a time t and at position j at time $t+1$. The transition matrix $\Pi = (\pi_{ij})$ is obtained by row-normalizing M so that $\pi_{ij} = \frac{m_{ij}}{\sum_l m_{il}}$. Hence, for all rows i, $\sum_j \pi_{ij} = 1$ and Π is a well-defined probability matrix for the transitions between pair of nodes in the network. Notice that for both M and Π the diagonal elements can be significantly different from zero and therefore the resulting networks display self-loops. Moreover, both matrices M and Π correspond to directed, weighted networks, and therefore can be thought as representing flows across the networks. For completeness, we also define the symmetrized versions of the matrices above, namely the symmetrized jump matrix $M^{\rm symm} = (M + M^{\rm T})/2$, Π and the symmetrized transition matrix $\Pi^{\rm symm} = (\Pi + \Pi^{\rm T})/2$. The corresponding weighted networks are undirected and represent a first coarse-graining of the information contained in the dynamical flows. It is thus interesting to compare these two to understand how much information is lost in the coarse-graining.

We performed community detection algorithms by optimizing modularity [9, 10]. To ensure consistency, we checked the results under different heuristics and repeated detections [11, 12]. Figure 4 shows the communities extracted from the M, $M^{\rm symm}$, Π and $\Pi^{\rm symm}$ matrices. The coloured hulls are included for comparison and indicate the Pardus cluster to which each sector belongs. For comparison, in figure 5 we plot the communities obtained from the topological quantities, namely the directed weighted adjacency matrix A, the undirected unweighted matrix $A^{\rm wh}$ and the inverse distance matrix $D^{\rm inv}$. One can easily see that the communities extracted from the transition matrices appear to reproduce much better the cluster structure as opposed to the topological communities.

Notice also that the partitions obtained for the dynamical transition matrices contain communities composed of a single node. Although unusual in community detection, this result is consistent with the mobility patterns. In fact, we measure the positions of players at the same time every day. Then, the presence of non zero values on the diagonal of M, $M^{\rm symm}$, Π and $\Pi^{\rm symm}$ simply means that there is a positive probability for a player to be found again on the same node after 24 hours, implying that the player either stayed still on the node or traveled but came back to its original position within 24 hours. These self-loops are responsible for the presence of single-node communities in the dynamical matrices and for their absence in the topological ones, where there are no self-loops.

We find a different number of communities for different matrices, making it hard to come to a conclusion regarding which one is the closest to the Pardus cluster structure. To quantify the relative goodness of the partitions obtained from the various matrices, we calculate three measures of clustering similarity: the Fowlkes and Mallows index $\mathcal{F}$ [13], the Rand's criterion $\mathcal{R}$ [14] and the normalized information variation (NVI) [15]. Consider a set of nodes $\mathcal{T}$ of cardinality n and two partitions $\mathcal{C}$ and $\mathcal{C}'$ of $\mathcal{T}$, then the set of all unordered pairs of elements of $\mathcal{T}$ is the union of the sets [16, 17]:

t_{11} is the set of pairs the same community under $\mathcal{C}$ and $\mathcal{C}'$;



Supplementary Figure 4. Extracting communities from mobility patterns. Communities found for (a) the jump matrix M, (b) the symmetrized jump matrix $M^{\text{symm}} = (M + M^{\text{T}})/2$, (c) the transition matrix Π and (d) the symmetrized transition matrix $\Pi^{\text{symm}} = (\Pi + \Pi^{\text{T}})/2$. Different node colours represent the different communities found, while the 20 different colour-shaded areas indicate the predefined socio-economic clusters as in Fig. 1 of the main text. The communities found through the information of motions reproduces well the bulk of the Pardus cluster structure, with a few exceptions along borders where some nodes are assigned to wrong clusters. The Fowlkes and Mallows index $\mathcal{F}$ is close to 1 for all detected partitions, reflecting the good match. For more measures, see Supplementary Table II.

t_{01} is the set of pairs not in the same community under $\mathcal{C}$ but under the same community in $\mathcal{C}'$;

t_{10} is the set of pairs in the same community under $\mathcal{C}$ but not under the same community in $\mathcal{C}'$;

t_{00} is the set of pairs not in the same community under $\mathcal{C}$ and $\mathcal{C}'$;

and n_{11}, n_{01}, n_{10}, n_{00} are their respective cardinalities (and $n_{11}+n_{01}+n_{10}+n_{00} = n(n-1)/2$). The $\mathcal{F}$ and $\mathcal{R}$ indices are then given by:

$$\mathcal{F} = \frac{n_{11}}{\sqrt{(n_{11}+n_{10})(n_{11}+n_{01})}} \qquad \mathcal{R} = \frac{2(n_{11}+n_{00})}{n(n-1)} \tag{1}$$

which are essentially two ways of quantifying how well the partitions match pairs of nodes. Therefore a perfect match between two partitions will have $\mathcal{F}, \mathcal{R} = 1$. The *Variation of Information* (VI) is a measure based on information

Matrix	Network Properties	n_{comm}	$\mathcal{F}$	$\mathcal{R}$	NVI
Clusters	—	20	1	1	0
A	directed, unweighted	18	0.678	0.963	0.179
A^{wh}	undirected unweighted	17	0.655	0.957	0.180
D^{inv}	directed, weighted	6	0.489	0.864	0.271
M	directed, weighted	23	0.963	0.996	0.025
M^{symm}	symmetrized, weighted	22	0.957	0.995	0.026
Π	directed, weighted	14	0.812	0.973	0.075
Π^{symm}	symmetrized, weighted	19	0.999	0.993	0.036

Supplementary Table II. Overview of community detection results for the studied matrices. From left to right, the columns correspond to: the studied matrix, the properties of the corresponding network, the number of communities found n_{comm}, the scores for the Fowlkes-Mallows index $\mathcal{F}$ [13], the adjusted Rand's criterion $\mathcal{R}$ [14] and, finally, the normalized information variation (NVI) [15]. For reference, the first row contains the values for the Pardus cluster structure. The closer the indices $\mathcal{F}$ and $\mathcal{R}$ are to 1, and the closer the NVI is to 0, the better a found partition matches the clusters. The values reported clearly indicate that the Pardus cluster structure is faithfully reproduced by the player mobility. On the other hand, the topological, non-dynamic properties (e.g. adjacency matrix, distance matrix) produce partitions that are very different from the Pardus cluster structure.

theoretical concepts and represents the informational distance between two partitions. Therefore, if the VI is large, the two partitions are very dissimilar. The VI of a partition is bounded by $\log_2 n$, hence it is possible to normalize it, obtaining the *Normalized Variation of Information* (NVI $\in (0,1)$):

$$\mathcal{NVI}(\mathcal{C},\mathcal{C}') = \frac{\mathcal{VI}(\mathcal{C},\mathcal{C}')}{\log_2 n} \tag{2}$$

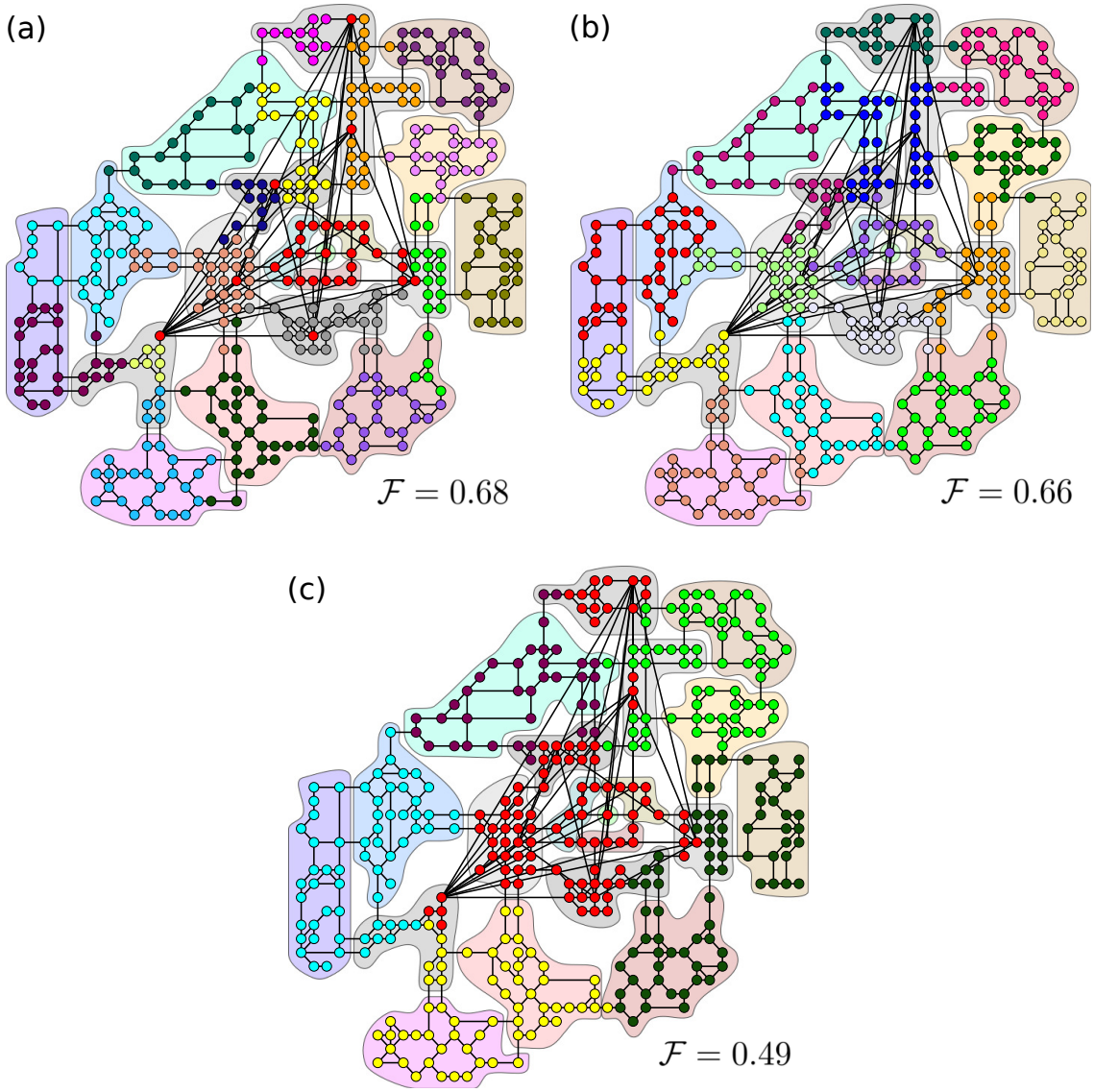
where

$$\mathcal{VI}(\mathcal{C},\mathcal{C}') = \mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}') - 2\mathcal{I}(\mathcal{C},\mathcal{C}') \tag{3}$$

The terms in equation (3) are the entropy $\mathcal{H}(\mathcal{C})$ of partition $\mathcal{C}$ and the mutual information between two partitions $\mathcal{C}$ and $\mathcal{C}'$ [17]:

$$\mathcal{H}(\mathcal{C}) = -\sum_{i=1}^{k} P(i)\log_2 P(i) \qquad \mathcal{I}(\mathcal{C},\mathcal{C}') = \sum_{i=1}^{k}\sum_{j=1}^{l} P(i,j)\log_2 \frac{P(i,j)}{P(i)P(j)} \tag{4}$$

where $P(i) = \frac{|C_i|}{n}$ is the probability that an element of $\mathcal{T}$ chosen at random belongs to community $C_i \in \mathcal{C}$, and $P(i,j) = \frac{|C_i \bigcap C'_j|}{n}$ the probability that an element belongs to $C_i \in \mathcal{C}$ and to $C'_j \in \mathcal{C}'$.
Supplementary Table II reports the values obtained for the studied matrices. The values of the Fowlkes-Mallows and Rand indices for the dynamical communities are much closer to 1 than the ones for the topological communities. The result is confirmed also by the NVI values, where we measured very small values for the dynamical partitions, indicating that player mobility follows closely the Pardus cluster structure. It could be argued that this similarity emerges from the topological structure of the network. However, we also found a difference of almost one order of magnitude between the dynamical and topological partitions and thus such hypothesis is not supported, that is the topological properties (e.g. adjacency matrix, distance matrix) produce partitions that are very different from the dynamical ones and the Pardus cluster one and cannot therefore be considered as the underlying mechanism of the mobility patterns. Moreover, this result is robust under different measures of player movement, as shown by the remarkable stability of the values of the clustering similarity measures for the other dynamical cases, M^{symm}, Π and Π^{symm}, which stay close to the ones obtained for M. Therefore, our conclusions cannot be considered an artifact of the particular measure we adopted.



Supplementary Figure 5. Extracting communities from topological information. Communities found for (a) the adjacency matrix A, (b) the adjacency matrix A^{wh} in which the X/Y-holes were removed yielding an undirected unweighted network, and (c) the distance matrix D^{inv}. Different node colours represent the different communities found, while the 20 different colour-shaded areas indicate the predefined socio-economic clusters as in Fig. 1 of the main text. The partitions obtained from the adjacency matrices produce communities that cross over the borders between clusters and therefore do not recover the clusters well. This is particularly evident in the case of D^{inv} where only 6 communities are found. The Fowlkes and Mallows index $\mathcal{F}$ is not close to 1 for all detected partitions, reflecting the bad match. For more measures, see Supplementary Table II.

[1] Szell, M. & Thurner, S. Measuring social dynamics in a massive multiplayer online game. *Social Networks* **32**, 313–329 (2010).

[2] Szell, M., Lambiotte, R. & Thurner, S. Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci. USA* **107**, 13636–13641 (2010).

[3] Castronova, E. *Synthetic Worlds: The Business and Culture of Online Games* (University of Chicago Press, Chicago, 2005).

[4] Song, C., Qu, Z., Blumm, N. & Barabási, A. Limits of predictability in human mobility. *Science* **327**, 1018 (2010).

[5] crowdflow.net.

[6] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. Complex networks: structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).

[7] Dorogovtsev, S. & Mendes, J. *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, 2003).

[8] Latora, V. & Marchiori, M. Efficient behavior of small-world networks. *Phys. Rev. Lett.* **87**, 198701 (2001).

[9] Arenas, A., Fernández, A. & Gómez, S. Analysis of the structure of complex networks at different resolution levels. *New J. of Phys.* **10**, 053039 (2008).

[10] Newman, M. Analysis of weighted networks. *Phys. Rev. E* **70**, 056131 (2004).

[11] Richardson, T., Mucha, P. & Porter, M. Spectral tripartitioning of networks. *Phys. Rev. E* **80** (2009).

[12] Kernighan, B. & Lin, S. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal* 291–307 (1970).

[13] Fowlkes, E. & Mallows, C. *A method for comparing two hierarchical algorithms*, vol. 78 (Journal of the American Statistical Association, 1983).

[14] Rand, W. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, pp. 846–850 (1971).

[15] Cover, T. & Thomas, J. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, 2006), 2 edn.

[16] Meila, M. Comparing clusterings–an information based distance. *Journal of Multivariate Analysis* **98**, 873 – 895 (2007).

[17] Brandes, U. *et al.* On finding graph clusterings with maximum modularity. In *Graph-Theoretic Concepts in Computer Science*, 121–132 (Springer, 2007).