Supplemental Material: U.S. Social Fragmentation at Multiple Scales

Leila Hedayatifar, Rachel A. Rigg, Yaneer Bar-Yam and Alfredo J. Morales*
New England Complex Systems Institute
277 Broadway, Cambridge, MA, 02139
* E-mail: alfredo@necsi.edu

S1.1 Contribution of states to communities in the mobility and communication networks

To further quantify how state borders contribute to communities, we counted the number of nodes for each community in each state and normalized to the size of the largest community. Figure S1 shows the contribution of states to communities for both the mobility and communication networks with resolution $\gamma = 1$. For many of the states, the borders of states match their communities, while other states have administrative borders that deviate from the ways people travel and communicate.

S1.2 Dissimilarity between communities in distance-based networks and real networks

The strength of communications between people has an inverse relationship with their distance from each other. Previous research has shown the relationship is inversely proportional to distance or distance squared [1, 2]. To investigate the effect of distance or distance squared on the formation of communities, we re-weighted the links in the mobility and communication networks based on the distance between any two connected locations.

Figure S2 shows the detected communities in both networks with links that are rearranged to be equal to the inverse of distance $1/d_{ij}$ between locations i and j. Panels (a) and (b) in this figure depict the communities, and panels (c) and (d) quantify the similarity of the communities. Although communities are geo-fragmented (36 communities in the mobility network and 30 communities in the communication network), they are not similar to the communities in the real networks. In fact, in some areas, communities almost overlap with each other, but in most areas, the artificial networks end up with smaller patches that do not match with any real community. Figure S3 shows the results for the artificial networks with links that are created based on $1/d_{ij}^2$ between two connected locations i and j. This dissimilarity between the communities in the distance-based networks and the real networks is evidence of the impact of other parameters beside distance.

S1.3 Stability of communities and overlapping communities

In Figure S4, we quantify the stability of detected communities and identify areas in which communities overlap with each other. In the Louvain method, communities refer to the regions in which

nodes are more connected to each other than the rest of the network. However, due to the possible local minima in the Louvain algorithm, some nodes may be detected as part of other communities in the next run [3, 4]. We generated an ensemble of 120 realizations to quantify how much the detected communities are stable across the realizations and in which areas communities overlap. In panels (a) and (b) of Figure S4, the color bar shows how stable a location is in a given community as a percentage of realizations. Panel (a) shows that in many areas, communities are 100% stable, and all the areas have at least 40% stability. In some areas, larger communities split into two communities. For example, Georgia (GA) and Alabama (AL) encompass one community in some realizations, while in others, AL splits off as a separate community. Similarly, Indiana (IN), Kentucky (KY) and Tennessee (TN) form a single community in some realizations, while in others, IN manifests as a separate community. In other cases, areas sometimes exhibit a small overlapping part between two stable communities, such as the small section in eastern New York State (NY) that sometimes appears as part of the red community in Figure 1. Connecticut (CN) is the state that has the largest instability. In Figure 1, it is part of the New England community (purple community), but in many realizations it appears as part of the New York City and surrounding area community (red community). In the communication network (panel (b)), some overlapping areas are the size of multiple states, demonstrating that sometimes a whole state appears in another community.

In panels (c) and (d) of Figure S4, we count the number of connections outside the community for each location. In panel (c), black spots are the locations that only have inside community connections, which tend to be suburban and rural areas around the cities. Red spots represent locations with more than 100 outside community connections, which are city centers. Roads are also clear in the figure, as people tweet on the road far from the locations where they spend most of their time. In panel (d), the number of black spots decreases while the red areas increase as compared to panel (c). These differences show the larger distances that people communicate with each other versus physical travel. Red spots in these two panels have a higher chance to appear as part of another community, especially ones that are on the border of two communities.

Panels (e) and (f) show the frequency of detected communities over the 120 realizations for both networks. For the mobility network, all realizations generated a range of $19-22$ communities, with more than 80% of realizations having 20 communities. For the communication network, the range was $14-17$ communities, with more than 80% of realizations having 15 communities. This shows the relative stability of the network, with only a few communities that were likely to split into two.

S1.4 Comparing communities detected by the modularity optimization and Infomap methods

In Figure S5, we show the communities detected by the Infomap method with communities detected by the generalized modulariy optimization method for (a) mobility and (b) communication networks. Compared to the modularity method in which communities represent connected areas that deviate the most from a null model [5], the Infomap method is based on the flow of information within the network, and communities represent the areas a random walker tends to stay in for a long time [6]. The algorithm of community detection with Infomap is very similar to the Louvain method [7, 6]. It starts by considering each node as a single module and joining the neighboring nodes into supermodules in a random sequential order. Nodes move to their neighboring modules

to reach the largest decrease in the description length given by a map equation. If the movement of a node to a neighboring module does not reduce its description length, it stays in its own module. This procedure is repeated by a new sequential order in each time step, until movements do not decrease the map equation. Movements in each level start from the formed modules in the previous level. This hierarchical procedure continues until reaching a minimum value for the map equation.

The Infomap method detects 512 communities in the mobility network (Figure S5-(a)) and 459 communities in the communication network (Figure S5-(b)). Most of the communities reflect cities and their suburbs. However, in the communication network, the communities are larger, with some in the size of the states. In panel (c) of Figure S5, we compare the similarity of detected communities for the Infomap and modularity methods. The x-axis shows resolution parameters for the modularity method. As a measure of similarity, the y-axis shows the average value of the three scores, Purity, Adjusted Rank and Fowlkes-Mallows Indexes. The highest similarity of communities in the mobility network is about 55 percent and occurs for resolution parameters of 20 to 40. Meanwhile, the similarity of communities in the communication network is larger at about 60 percent around resolution 8 but then decreases sharply with increasing resolution parameters.

S1.5 Structural similarities of the mobility and communication networks

We compare structural properties of the mobility and communication networks by means of centrality measures, edge weights and multi-scale structure. The structural properties of both networks are consistent with each other, showing an interplay between how people explore the physical space and communicate on Twitter.

In Figure S6 (a), we show a scatter plot of degree centrality for each location in both networks at $\gamma = 1$, colored by their eigenvector centrality in logarithmic scale. While most locations are poorly connected, a few of them have an extremely high degree, corresponding to densely populated areas in large cities. Locations with a higher degree centrality in both mobility and communication networks also have high eigenvector centrality, which means that these locations are central relative to where information flows.

Next, we compare edge weight and length for both networks in Figure S6 (b). The edge length is estimated as the geographical distance between the locations' centroids. Edges that have high weights also have small lengths, reflecting daily, short distance travels seen in cities and localized communication.

Finally, we compare the multi-scale structure of the network fragmentation. Social fragmentation can be seen at multiple scales using the generalized modularity optimization method. This method seeks partitions at various scales by considering the resolution parameter γ. We compare the two networks at roughly similar partition sizes (mobility and communication, respectively, at γ=0.1 and 0.3, γ=0.2 and 0.6, γ=0.6 and 0.9, or γ=0.7 and 1.0). These pairs of γ values represent partitions in both networks with the highest similarities (see Figure 6 in the main text). In Figure S7, we show the partitions comparison using an alluvial diagram for each set of γ values (panels). The alluvial diagrams map the corresponding number of nodes at each module of the mobility network (left axis) onto each module of the communication network (right axis).

S1.6 Comparing t-SNE analyisis in real dataset with the randomized dataset

To validate the clusters we see in the t-SNE analysis of the real data, we performed t-SNE on a randomized dataset (see Figure S8). In the randomized dataset, we kept the number of instances of each hashtag the same but randomized the location for each hashtag use. We compared the distribution of the distances between locations per community in the t-SNE space for the real dataset to those of the randomized dataset. The patterns we see in the t-SNE of the real dataset were statistically different from those of the randomized dataset for all communities ($p < 0.001$).

S1.7 Additional results from varying the model parameters

In Figure S9, we show the effect of changing the spatial growth term ν in the network model. This term gives preference to locations in which the average degree of nearest neighbors is higher. We set the α and β exponents at 0.5 and 1.5, respectively, which generates geographical fragmentation patterns even with $\nu = 0$ (left panels). As shown in Figure S9 (a), increasing the ν exponent (from left to right panels) concentrates the connections around hotspots, recreating the growth of cities. However, Figure S9 (b) and (c) show that while small values of ν increase the number of communities and their modularity, large values of ν lead to less cohesive borders between communities and a decrease in modularity.

Next, we investigated how changing all three parameters in the model causes deviations between the model and the real data. We used the Kolmogorov-Smirnov (K-S) test, a measure of similarity between two distributions, to determine the similarity of the network degree distributions, which are a measure of network connectivity. Degree distributions for the model were calculated by averaging over 20 realizations for each set of parameters. Figure S10 represents the similarity of degree distributions of locations for the model versus the empirical mobility network. Moving from the top left panel to the bottom right panel, we see that the degree distributions for the model and the mobility network data deviate from each other as the ν exponent increases.

We also determined how changing the model parameters affects the modularity of the network. Figure S11 shows the average modularity from 20 realizations for different values of the three exponents. Overall, increasing the strength of the preferential attachment process (controlled by α) and the spatial growth process (controlled by ν) destroys geographical patches and reduces modularity. Conversely, the human mobility gravity process (controlled by β) is the only exponent that increases these fragmented geographical patches. Note that modularity is 0.83 for the mobility network from the Twitter data.

References

[1] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623–11628, 2005.

[2] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. *In Proceedings of WWW*, 10:61–70, 2010. doi: 10.1145/1772690.1772698.

[3] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. Nunes Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104 (39):15224–15229, 2007.

[4] B. H. Good, Y-A. de Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81:046106, 2010.

[5] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69 2 Pt 2:026113, 2004.

[6] L. Bohlin, D. Edler, A. Lancichinetti, and M. Rosvall. *Community Detection and Visualization of Networks with the Map Equation Framework. In: Ding Y., Rousseau R., Wolfram D. (eds) Measuring Scholarly Impact.* Springer, Cham, 2014.

[7] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

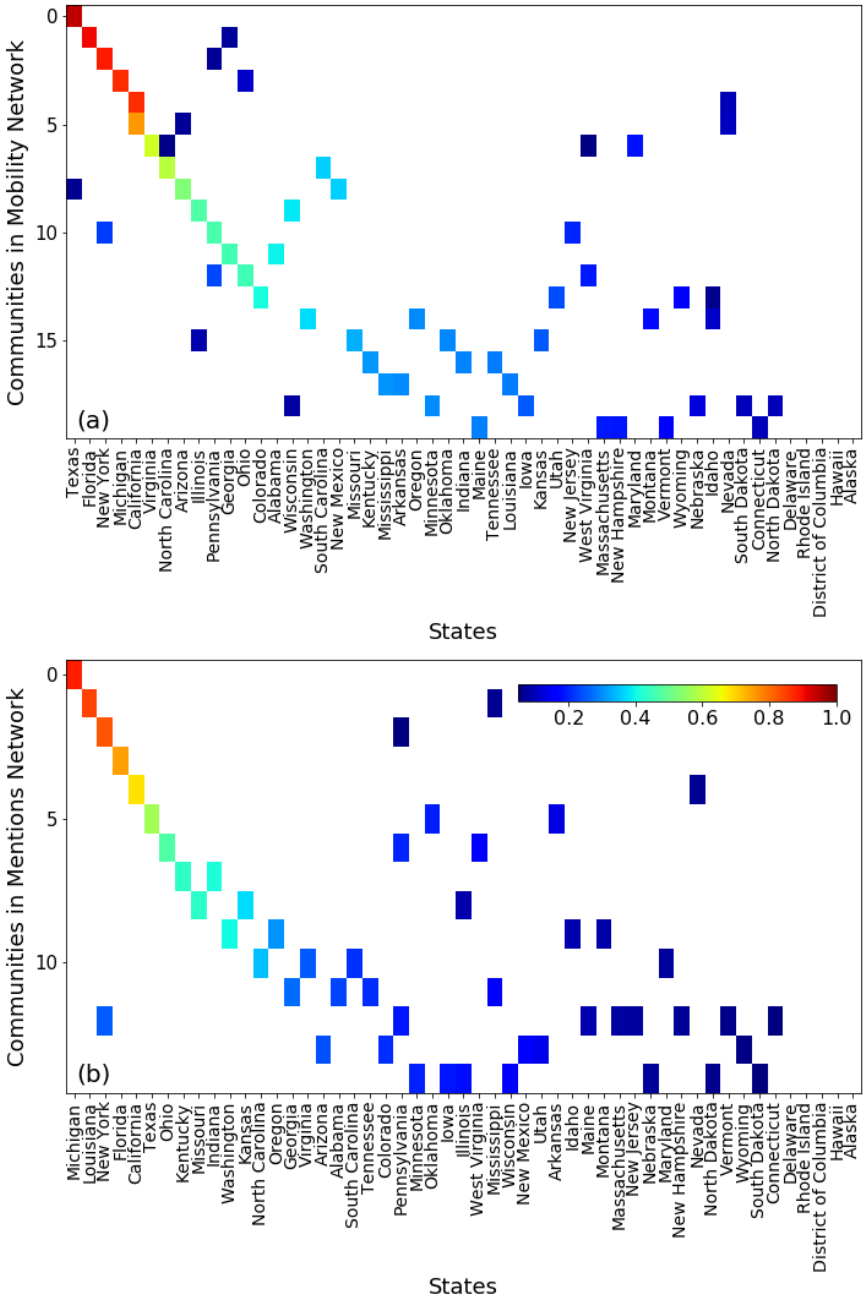


Figure S1: Contribution of states to communities. Graphs depict the number of communities detected (y-axis) in each state (x-axis) for (a) mobility and (b) communication networks generated with $\gamma = 1$. Colors (scale inset, panel (b)) indicate overlap between state and community boundaries, with red indicating greatest overlap.

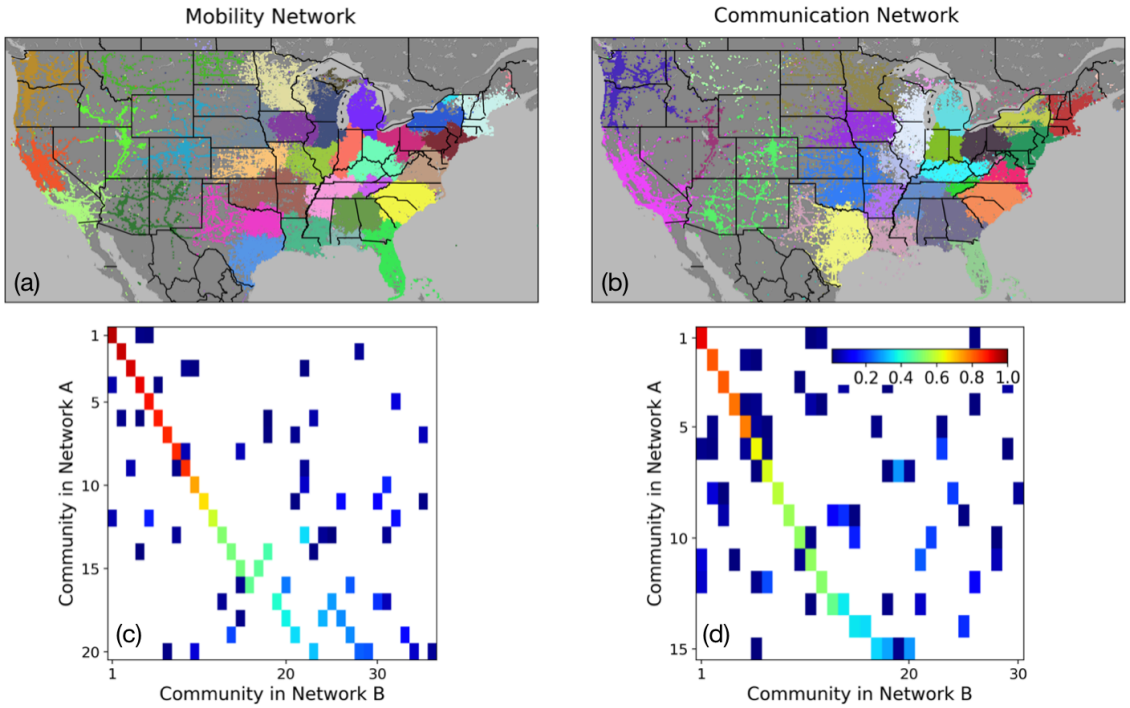


Figure S2: Dissimilarity between communities in real networks (mobility and communication) and artificial ones in which the weight of the links represents the inverse distance between the connected locations i and j, $1/d_{ij}$. Panels (a) and (b) show the detected communities in the artificial mobility and communication networks, respectively. Panels (c) and (d) show the overlap of communities between real and artificial networks of mobility and communication, respectively. The x-axes show the communities in the real networks (labeled "Network B"), and the y-axes show the communities in the artificial networks (labeled "Network A"). Communities are ordered by decreasing overlap. Cell colors represent the number of nodes overlapping between the two networks in each community, normalized by the size of the communities per row (scale inset), with no overlap indicated in white. Although the communities from artificial networks are geo-fragmented, they do not match the communities detected in real networks.

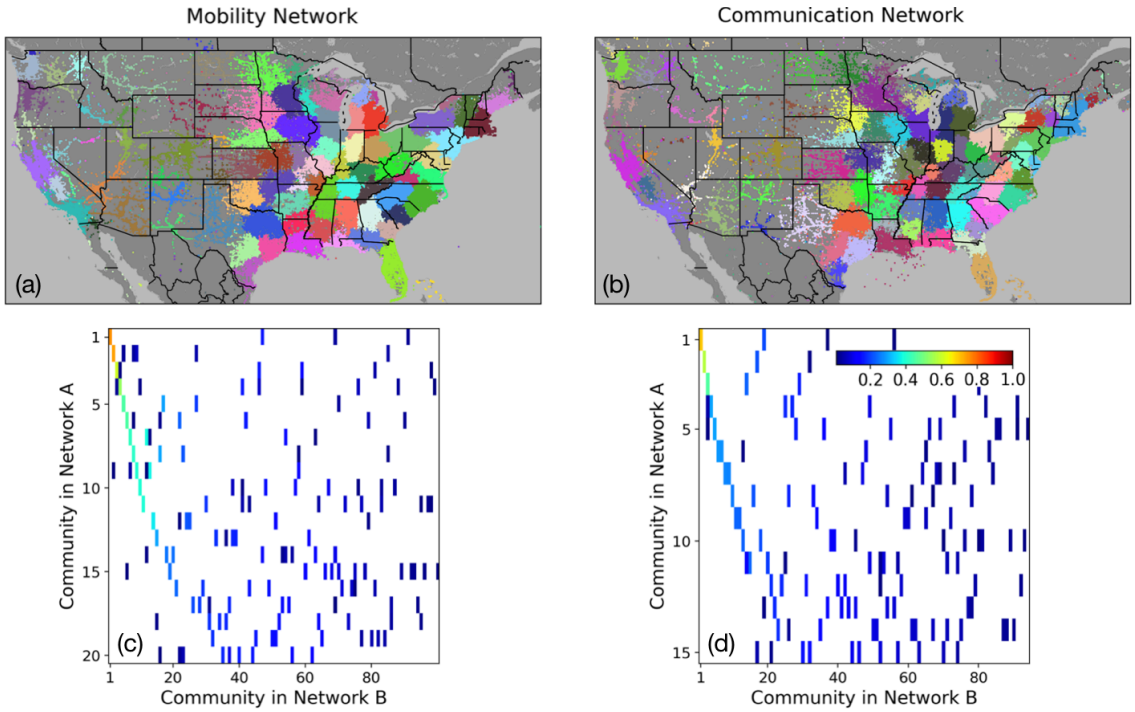


Figure S3: Dissimilarity between communities in real networks (mobility and communication) and artificial ones in which the weight of the links represents the inverse square distance between the connected locations i and j, $1/d_{ij}^2$. Panels (a) and (b) show the detected communities in the artificial mobility and communication networks, respectively. Panels (c) and (d) show the overlap of communities between real and artificial networks of mobility and communication, respectively. The x-axes show the communities in the real networks (labeled "Network B"), and the y-axes show the communities in the artificial networks (labeled "Network A"). Communities are ordered by decreasing level of overlap. Cell colors represent the number of nodes overlapping between the two networks in each community, normalized by the size of the communities per row (scale inset), with no overlap indicated in white. Although the communities from artificial networks are geo-fragmented, they do not match the communities detected in real networks.

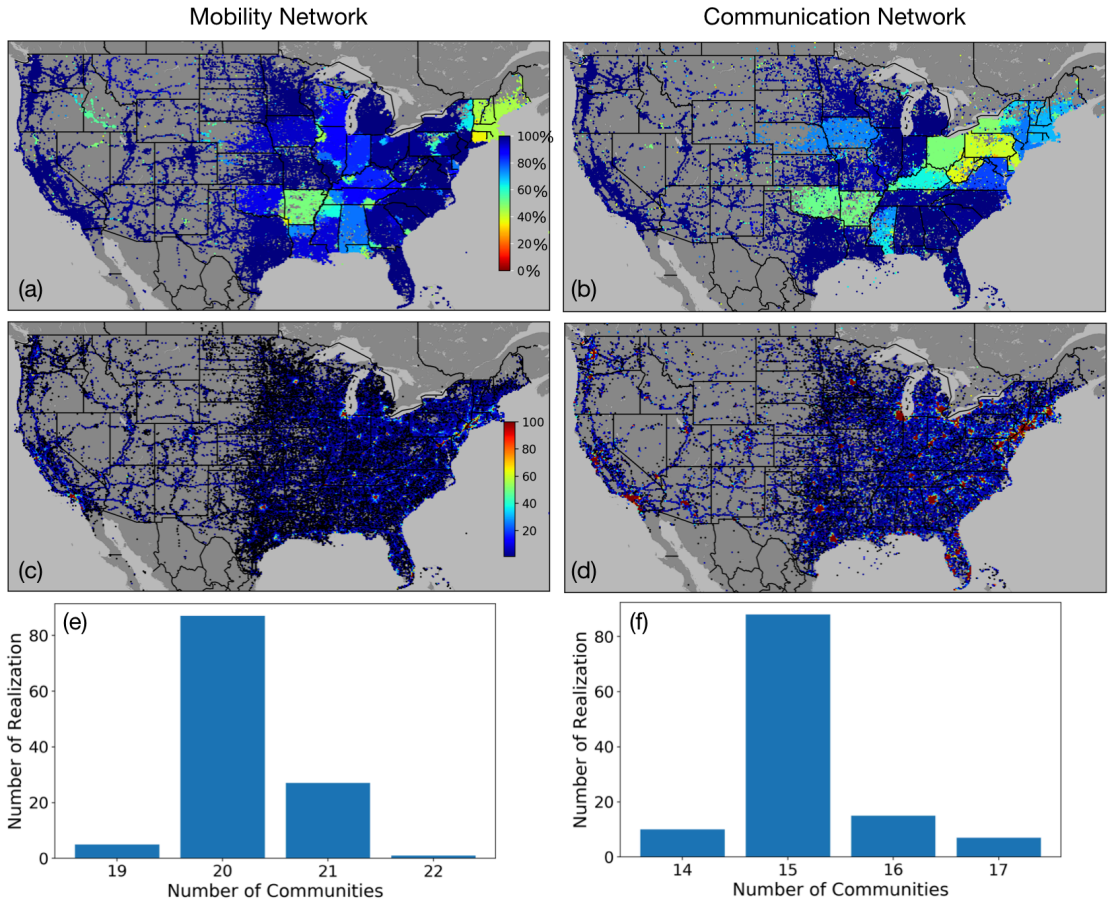


Figure S4: Stability and overlap of detected communities in mobility and communication networks. We created 120 realizations for detection of communities for each network. Panels (a) and (b) show the the stability of locations to their communities. The color bar (inset) represents the percentage at which a location appears in the same community over all realizations; blue areas are the most stable communities. Panels (c) and (d) show how many connections a location makes with locations outside its community in the mobility and communication networks, respectively. The color bar (inset) indicates the number of connections with outside areas, with red areas having more than 100 outside connections and black areas having no outside connections. Panels (e) and (f) show the frequency of detected communities for all realizations; 20 and 15 communities are the most frequent number of detected communities in the mobility and communication networks, respectively.

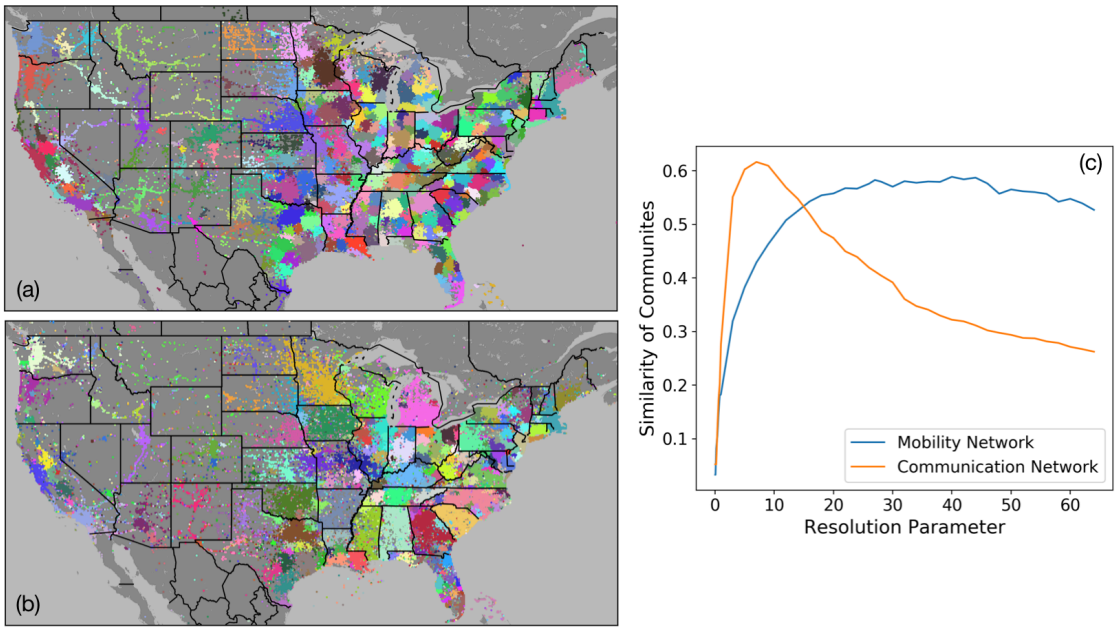


Figure S5: Detected communities in (a) mobility and (b) communication networks using the Infomap method. Applying the Infomap method gives smaller communities that mostly represent the city areas. Some detected communities in the communication network represent states. Panel (c) shows the similarity of detected communities using the Infomap method with the communities detected using the modularity optimization method (y-axis) at different resolution parameters (x-axis). To measure similarity of the communities, we used the average value of the three scores, Purity, Adjusted Rank, and Fowlkes-Mallows Indexes. Detected communities in the mobility network have more that 55 percent intersection over the resolution parameters 20-40. In the communication network, detected communities have the most similarity at resolution 8.

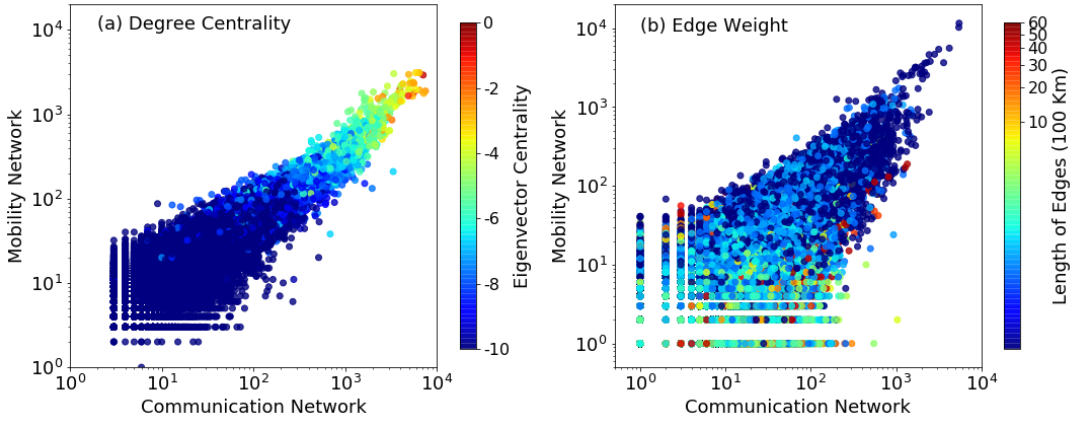
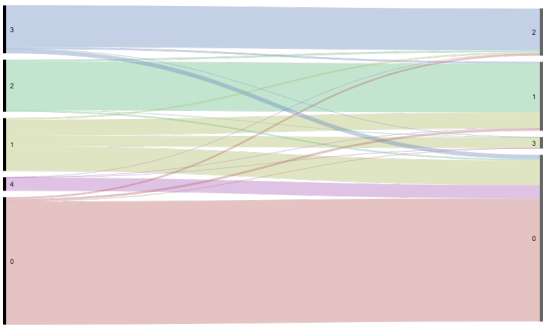
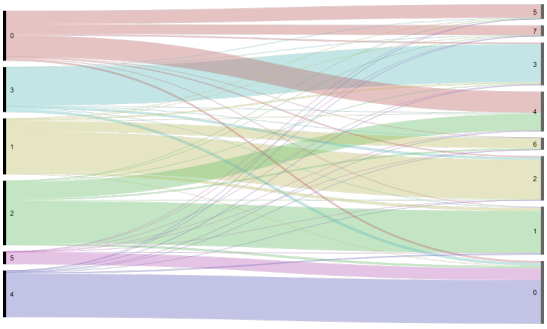


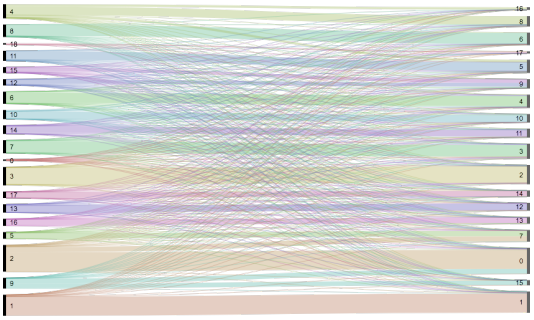
Figure S6: Similarity of network structure between the mobility and communication networks at $\gamma = 1$. Panel (a), scatter plot of degree centrality for each location in the mobility (y-axis) and communication (x-axis) networks, colored by the corresponding eigenvector centrality (scale on right). Panel (b), scatter plot of the edge weights for each location in the mobility (y-axis) and communication (x-axis) networks, colored by the edge length or distance between nodes (scale on right). Axes and color bars are in logarithmic scale.



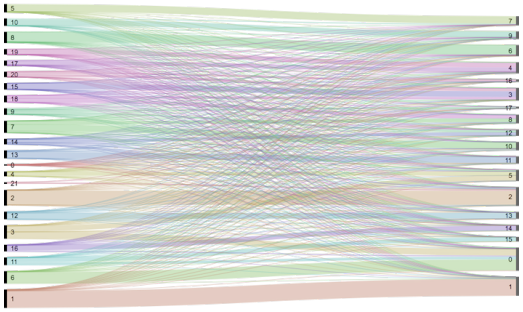
Mobility γ=0.1 - Communication γ=0.3



Mobility γ=0.2 - Communication γ=0.6



Mobility γ=0.6 - Communication γ=0.9



Mobility γ=0.7 - Communication γ=1.0

Figure S7: Comparison of the mobility and communication networks at γ values with highest similarity. In each panel, an alluvial diagram maps similarities of detected communities between the mobility (left axis) and the communication (right axis) networks. Values of γ are chosen from the four darkest red cells in Figure 6, which yield the highest similarity between the two networks.

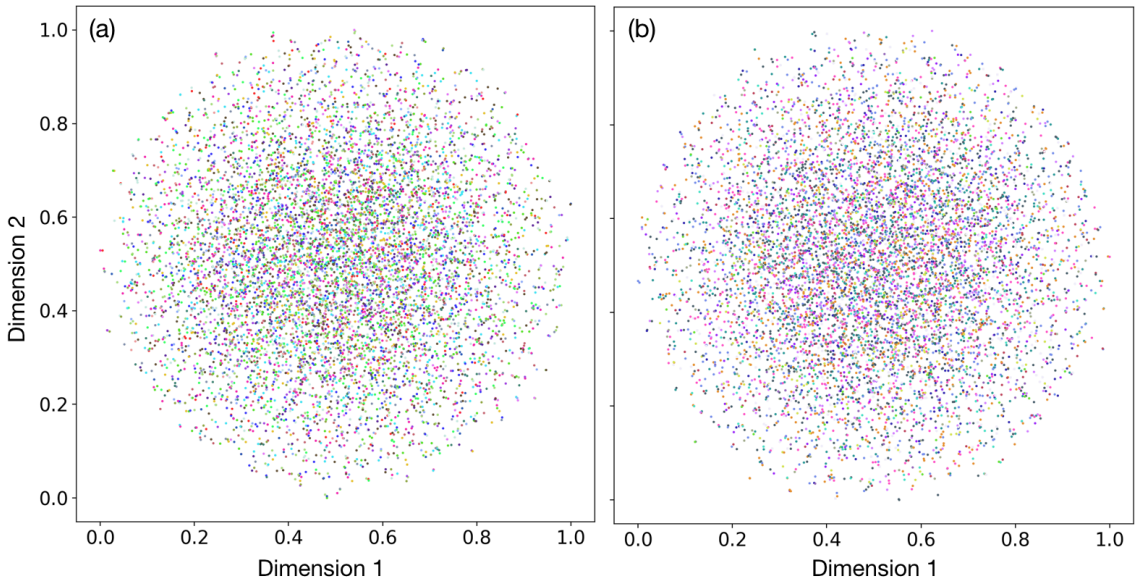


Figure S8: t-SNE analysis of hashtag use for a randomized dataset. Panel (a) shows the results of t-SNE analysis on the first 100 components of PCA analysis of hashtags in locations of the mobility network. Panel (b) shows the corresponding t-SNE analysis for the communication network. Instances of hashtag use have been preserved from the original datasets, but the locations have been randomized. Colors match those of the communities in Figures 1 and 2, panel (b).

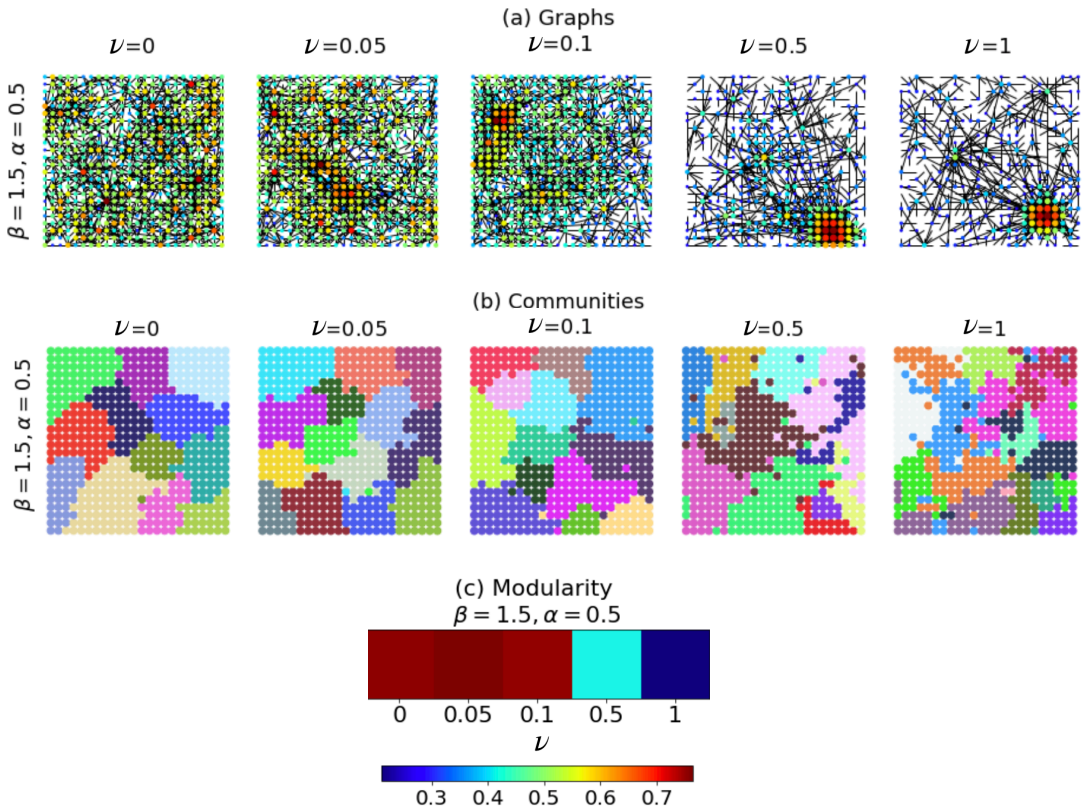


Figure S9: Degree centrality, fragmentation, and modularity for simulations with varied values of the resolution parameter ν (spatial growth). The value of ν is varied from $0, 0.05, 0.1, 0.5, 1$ (left to right), while $\alpha = 0.5$, $\beta = 1.5$, and the size of the lattice is 576 locations. (a) Spatial degree centrality. Nodes are colored by their degree centrality (from blue to red), and edges are plotted in black. (b) Spatial patches, shown in varying colors. (c) Modularity as a function of ν, indicated by color (scale below figure). Overall, increasing ν from 0 to 0.05 increases the connections between hotspots and the modularity of communities, but as ν increases further to 0.5 or 1, the modularity decreases and borders between communities become less clear.

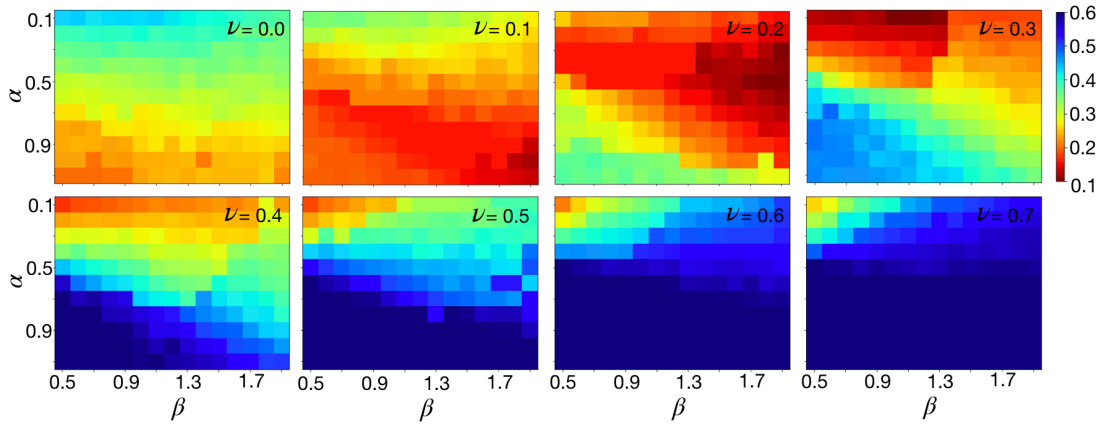


Figure S10: Similarity of degree distributions between simulated and real mobility networks. Matrices show different values of the parameters α (y-axis), β (x-axis) and ν (panels, upper left to lower right). Kolmogorov-Smirnov (K-S) scores are depicted with colors (scale on right), with the lowest K-S values in red, indicating that the degree distributions are similar between simulated and real mobility networks.

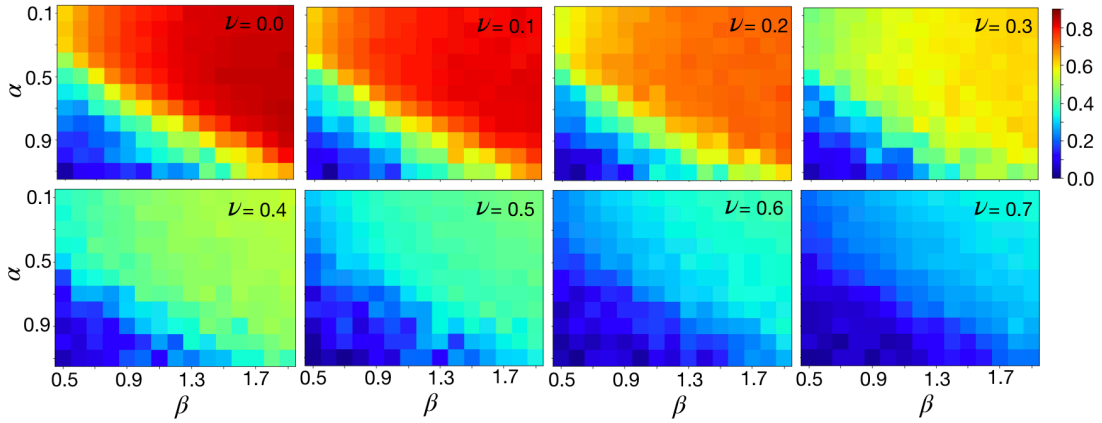


Figure S11: Modularity of detected communities in simulations with varying model parameters. Matrices show different values of the parameters α (y-axis), β (x-axis) and ν (panels, upper left to lower right). Modularity values are depicted with colors (scale on right), with the highest values in red at ~ 0.9. Note that modularity for the mobility network is 0.83.