Modeling the Evolution of the Global Migration Network

Stephanie Chen (schen751)

December 10, 2017

1 Introduction

Human migration has shaped the world since humans first came into being as a species; with recent rapid advancement of transportation and communications technology, however, migration has accelerated and expanded in scope to produce a truly global network, shaping politics, economies, and social demographics worldwide. By most predictions, migration will only continue to increase in rate, volume, and complexity in the coming years and decades.

Until recently, studies of human migration were limited to mostly bilateral or small-group multilateral cases between subsets of countries or regions, due to the lack of reliable global-scale migration data. Recent data releases by the United Nations and the World Bank, however, have opened the possibility of studying migration as an actual global network structure. Previous work on these sets has mostly focused on analyzing and modeling the static features of this network at different points in time. In this project, we investigate existing static migration models as well as temporal evolution models based on other networks, and we use these to develop a model representing the growth and evolution of the global migration network over time.

2 Literature review

2.1 Fagiolo & Mastrorillo, 2013

Fagiolo and Mastrorillo[1] examine the structure of a weighted, directed network based on the Global Migration Database (GMD), consisting of origin-destination bilateral counts of migrant stocks for 231 countries from 1960-2000.

They find that over the given period, network density increases, average path length decreases, and community structures merge so that the number of communities decreases. In terms of static properties, they find that edge weights are power-law distributed, dissortativity on both node degree and strength, and a high level of average clustering, for a summary view of "a relatively poorly concentrated network, with a binary small-world structure where a few hubs and a lot of local structures coexist." They compare the network to a null model (based on [4] as cited in [5]) where only node degree and strength are fixed, finding that local structure is insufficient to account for the observed properties.

They create instead a gravity model (Eq. 1, see [1] for variable interpretations) where weights are generated based on unary and binary variables like country population, contiguity, relative GDP, etc. and find that this model does reproduce the network's higher-level topology.

$$m_{ij} = P_i^{\alpha_1} P_j^{\alpha_2} d_{ij}^{\alpha_3} r Y_{ij}^{\alpha_4} \exp(\beta_i + \beta_j + \gamma Z_{ij}) \eta_{ij} \quad (1)$$

This model is heavily dependent on real-world socioeconomic factors such as GDP, language, and religion, which fits well to the global migration network itself but is not extensible to other networks with potentially similar underlying features. On the other hand, networks that are unrelated yet similar to this global-scale graph may not realistically exist, and [1] does include frequent points of comparison to the international trade network.

2.2 Porat & Benguigui, 2016

Porat and Benguigui[3] examine a migration flow — as opposed to migration stock — network between 2006-2010, finding in the degree distribution of the undirected network two bell-shaped groupings, one for small-degree nodes (about 80% of the network) and one for large-degree nodes (about 20%). They measure degree correlation, betweenness/closeness/eigenvector centrality, and clustering coefficient, finding dissortativity (similar to in [1]), a relationship between betweenness and clustering coefficient, and a clear separation between

small- and large-degree nodes in all measures, suggesting a strongly connected hub topology in the large-degree nodes and a dual-connection topology of edges to hubs and to local neighbors in the small-degree nodes.

The authors use a "two-group model" (Alg. 1) to model this relationship, where they generate random edges and then add additional edges to 20% "hub" nodes, finding that this model captures the observed properties, concluding that the large- and small-degree separation is the primary controller of the observed properties. Some "hub" properties are seen in [1] as well, but this model does not take into account edge direction or edge weight — both of which are significant features for real-world implications.

2.3 Leskovec, Kleinberg, & Faloutsos, 2007

Leskovec, Kleinberg, and Faloutsos[2] study 12 datasets of networks evolving over time, finding that contrary to conventional assumption that average degree stays constant and that network diameter grows slowly over time, real-world graphs exhibit densification and shrinking diameter over time as they evolve.

They propose a densification power law (Eq. 2) relating the graph's edge and node counts over time, as well as two generative models.

$$e(t) \propto n(t)^{\alpha} \tag{2}$$

The better of the two models, the "forest fire" model does this: new nodes link to "ambassador nodes" in the network and then recursively form links to those nodes' neighbors with some probability, thus "burning" through the network. This model exhibits both the densification power law and shrinking graph diameter, as well as heavy-tailed in- and out-degree distributions.

The forest fire model is only somewhat applicable as-is to this specific case: since the number of countries or sources of migration in the world is relatively constant, new nodes are effectively never added to the migration network, and the densification and shrinking-diameter properties are somewhat inevitable. However, because it is useful as a model of temporal evolution, we do use the forest fire model for comparison and make adjustments to the migration network, discussed below.

3 Method

3.1 Data

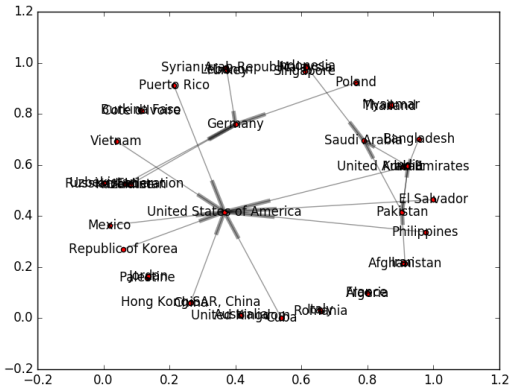

Figure 1: Subset of GMD, with only nodes with strength (total in- and outflow) greater than 1000000 shown.

One of the primary issues with migration network is the absence of a single comprehensive, definitive dataset; migration data is usually aggregated from census and immigration statistics kept per country, and even the aggregation process differs based on organization. We focused on the World Bank 1960-2000 set (GMD), which covers bilateral migration stocks every 10 years from 1960 to 2000.

We also examined the U.N.'s Global Migration Stock set (GMS), which covers bilateral stocks every 5 years from 1990-2015, but its data is more restricted, as it relies on official immigration/emigration statistics kept by each country. The GMD, on the other hand, is derived from a combination of national immigration statistics as well as census data and population surveys and is the basis for most of the existing literature, so we chose to focus there. The two sets also differ slightly in countries represented, due to political changes (ex. countries being split), disputed territories, and consideration of small island nations, but the countries involved are relatively small migration participants and did not contribute meaningfully to differences between the two graphs.

In our proposed model (discussed later), we also use national GDP (in today's U.S. dollars) and total population time-series data from the U.N., and geographical border data from the CIA World Factbook.

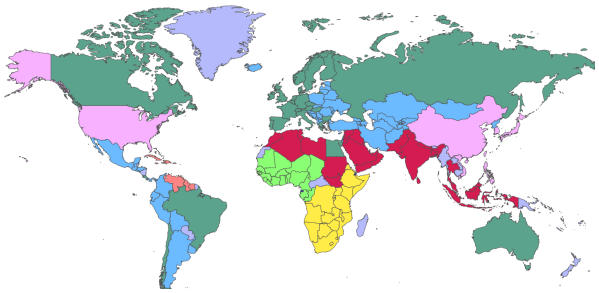

Figure 2: Spectral partition of the GMD in 2000 into 8 groups; we can see economic and cultural as well as geographical factors in play. Placement of the U.S., however, is fairly meaningless, as it is by far the largest "hub" in the graph and switches groups depending on cluster count.

3.2 Graph properties

The GMD has 231 nodes, with edge count increasing from about 16000 to nearly 24000 over the 40-year period. The graph densifies over time, as expected from [1], from 0.31 to 0.45. Average clustering coefficient (on the undirected graph) increases slightly over time, from 0.748 to 0.791. This overall high clustering level confirms previous work; we performed a spectral partition and found that clusters in the graph correspond to geographic regions as well as economic and cultural relationships (Fig. 2). Average shortest path length is extremely short and decreases over time as expected, from 1.54 in 1960 to 1.38 in 2000. The GMD is also weakly dissortative, again confirming previous work, with average degree assortativity over the time period at -0.200.

The GMS has 232 nodes, with edge count increasing from roughly 10000 to 11000 over the 25-year period. While most GMS statistics roughly matched the GMD, we initially found that the degree distributions were extremely different (Fig. 3).

We managed to account for most of this difference by setting a weight threshold of 10 for an edge to be included in the graph; the resulting filtered GMD is much more similar to the GMS in both degree and strength distribution. This makes sense, as the GMS's official statistics are less likely to account for ultra-small immigration and emigration levels.

Because other GMS statistics roughly match the GMD but show much less change over time, likely due to the majority of globalization having already happened during the period covered by the GMD, we focus on the GMD going forward. Consistent with previous work, we found a power-law relationship in node strength (Fig. 5, with average strength increasing over time.

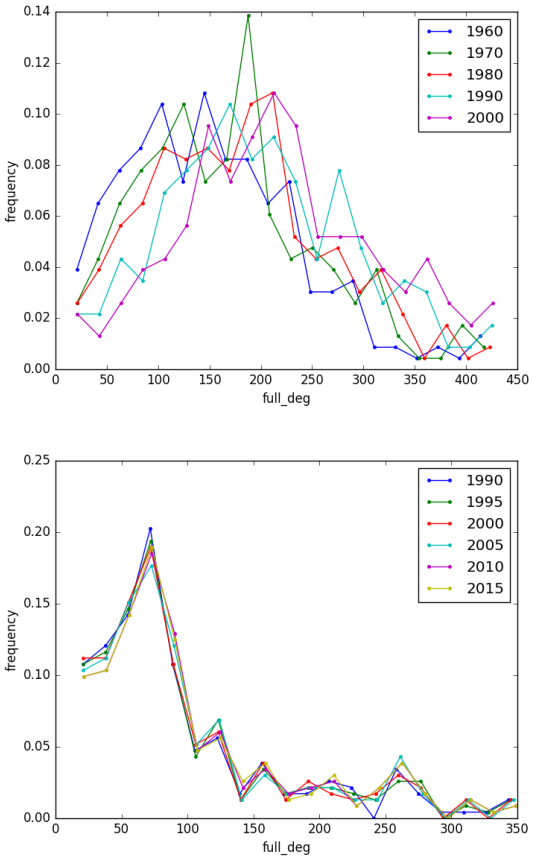

Figure 3: The extreme initial difference between the degree distributions of the GMD and GMS.

Going forward, in our modeling, we focused on accurately modeling these critical features presented above: high clustering, high density, dissortativity, short average path length, and a power-law strength distribution. As noted in [1], the combination of these features is fairly unusual, as it shows the GMD does not resemble other global networks, most of which show more of a preferential attachment/rich-get-richer structure, with fewer clusters and higher assortativity.

3.3 Null models

Two-hub model.
We first implemented the two-hub model (Alg. 1) from [3] due to its simplicity as well as an initial expectation that the two-hub structure made sense for the stock network as well as the flow network. We

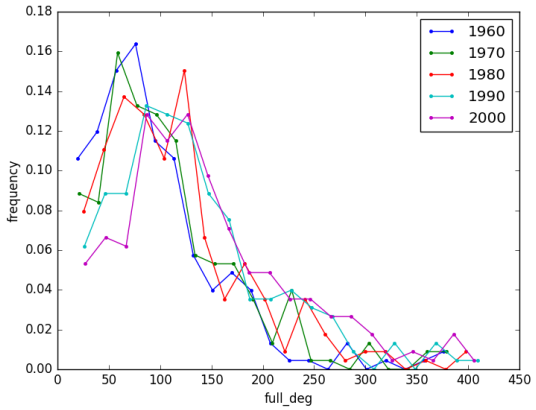

Figure 4: The GMD with a threshold weight filter of 10 resembles the GMS.

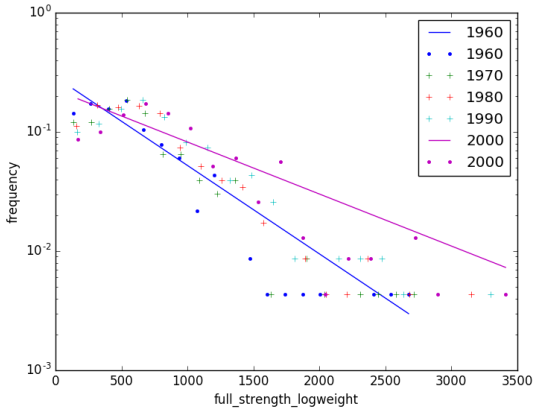

Figure 5: Plot of the node strength distribution. Strength is calculated here as the sum of log weights, following the example of [1], so this is not a strict log-log plot. Log fits are shown for 1960 and 2000, showing an overall increase in strength over time.

found, in fact, that most top-level properties of the network were fairly accurately represented by the two-hub model, which in an average of 5 runs produced a network of 232 nodes and 13536 edges, with slightly lower clustering coefficient, slightly higher density, and similar dissortativity and average path length.

However, we can see from the degree and strength distributions above that there is no clear two-hub structure in the GMD.

Algorithm 1 Two-hub model

procedure TWO-HUB(n)
  $g \leftarrow$ empty undirected graph
  for $i = 0; i < n; i++$ **do**
    add node i to g
  end for
  $hubs \leftarrow 20\% * n$ nodes in g.nodes
  for (u, v) with $u \in g.nodes, v \in g.nodes$ **do**
    if $rand() < 30\%$ **then**
      add edge (u, v) to g
    end if
  end for
  for $h \in hubs$ **do**
    create edges from h to $80\% * n$ in $g.nodes$
  end for
end procedure

Configuration model.
Instead, as our static null model, we used a standard configuration model based on the GMD's degree distribution from 1960. While the degree distribution matches, as expected, the configuration model was much less dense, less clustered, and less dissortative, confirming that the migration network's features are due to more than just local degree structure.

Forest fire model.
As a null model for "growing" the GMD graph, we use the forest fire model from [2], reproduced here (Alg. 2). To adapt the GMD, in which new nodes rarely join if at all, to the new node-based forest fire model, we set a threshold of degree of at least 200 (sum of both in- and out-degree) for entry into the network. We used a degree threshold after finding that weight (migration volume) thresholds disproportionately punished low-population countries, while migration volume as a percentage of population was unpredictable, increasing in some countries, decreasing in some, and fluctuating apparently arbitrarily in others. We then chose the 200-degree threshold after finding that it produced the greatest overall change across the network in terms of nodes and edges added (Fig. 6, while also following the densification power law from [2], with $\alpha = 1.86$.

This is a relatively high densification exponent even after attempting to account for the finite total nodes, likely due to the limitations of an objectively small overall network and only having 5 points of reference in time for the growth of the network. We find a best-fit forest fire model of 200 nodes with forward burning probability $p_f = 0.85$ and backward burning probability $p_b = 0.32$, seen in Fig. 7.

Model	Nodes	Edges	Density	Clustering coeff.	Assortativity	Path length
GMD (1960)	231	16485	0.310	0.748	-0.205	1.539
GMD (2000)	231	23718	0.446	0.791	-0.209	1.379
Two-hub	232	13536	0.505	0.655	-0.199	1.498
Config.	231	12490	0.235	0.535	-0.036	1.592
Forest fire	200	16874	0.424	0.939	-0.488	1.152

Table 1: Top-level graph statistics for the GMD and the null models tested.

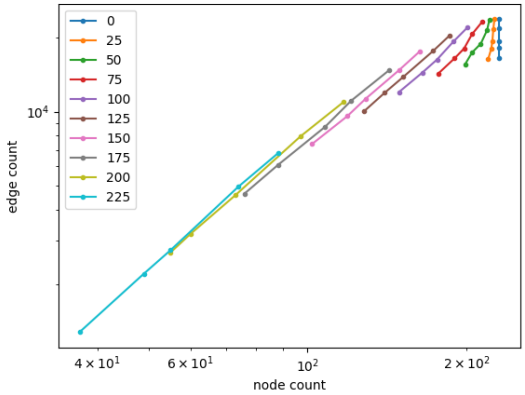

Figure 6: Log-log plot of change in the GMD network over the 1960-2000 period based on different degree cut-offs for entry into the network.

Algorithm 2 Forest fire model

procedure FOREST FIRE(n, p_f, p_b)
  $g \leftarrow$ single-node directed graph
  while $g.nodes < n$ **do**
    $v \leftarrow$ new node
    $w_0 \leftarrow$ random from $g.nodes$
    $q \leftarrow$ queue containing w_0
    $s \leftarrow$ empty set
    while q not empty **do**
      $w \leftarrow q.pop()$
      if $w \in s$ **then continue**
      add w to s, edge (v, w) to g
      $x \leftarrow \text{geo}(1 - p_f) - 1$
      $y \leftarrow \text{geo}(1 - p_b) - 1$
      for $w_i \in x$ outlinks of w **do**
        add edge (v, w_i) to g
        add w_i to q
      end for
      for $w_j \in y$ inlinks of w **do**
        add edge (v, w_j) to g
        add w_j to q
      end for
    end while
  end while
end procedure

This model has 200 nodes and 16874 edges, which is similar to the GMD, but is denser, has an extremely high clustering coefficient of 0.94, and shows an extremely different degree distribution (Fig. 8) representative of the high value of α.

Ultimately, the forest fire model is not well suited to the global migration network — [2] focuses on the growth of very large networks, with orders of magnitudes more nodes than the GMD. In addition, the forest fire model works on a fundamental assumption that only new nodes can point towards old nodes; nodes present at time t_i cannot form any outlinks to nodes at time $t_{i+1}, ...$ Though this makes sense for ex. citation networks, in the global migration network, an increase in emigration from a country doesn't prevent (and actually probably promotes) an increase in immigration to that country. Nevertheless, the forest fire model provides a good baseline model of network evolution against which we can compare.

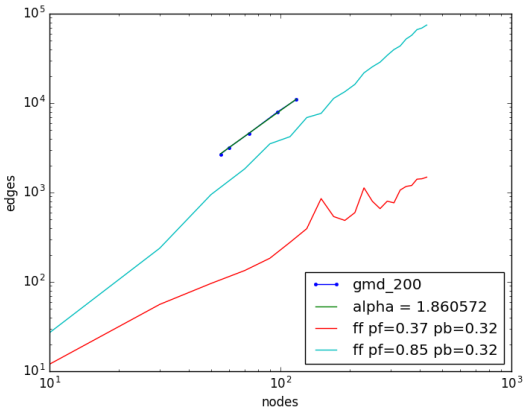

Figure 7: Edge-node plot of the GMD with a degree-200 threshold, against a best-fit forest fire model with $p_f = 0.85$, as well as the "realistic case" model from [2] with $p_f = 0.37$.

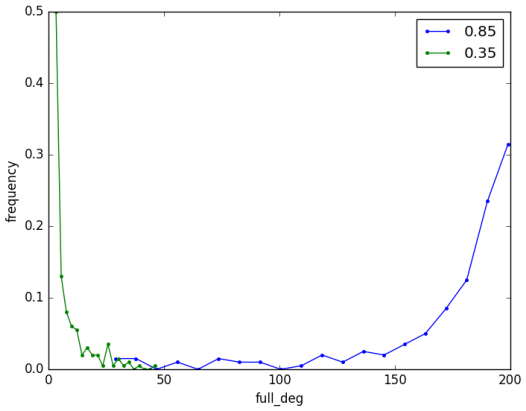

Figure 8: Degree distribution of the forest fire model with $p_f = 0.85$ and $p_b = 0.32$, against the "realistic" model with $p_f = 0.35$ for comparison.

3.4 Proposed model

We've seen from the above null models that the properties of the global migration graph are difficult to replicate without information in the network's geographical reality. We thus propose a model based on the community evolution structure commonly seen in social networks and grounded in simple geopolitical data. Though we initially proposed creating a general model unattached to real-world political and economic data (as in 1), we realized that it was impossible to model the effect of countries like the U.S. and U.K., which have minimal borders but enormous economic and political influence otherwise. In this model, we have tried to minimize the complexity of these real-world factors to avoid something similar to 1.

In our model (Alg. 3), countries begin with symmetric, high-weight edges to countries with which they share land borders; on each iteration, they "update" a randomly selected outlink to one of these bordering countries, or with some probability p, they "update" an outlink to a country randomly selected from a distribution weighted by GDP and population. Here, "updating" means creating an outlink if an edge doesn't already exist, and multiplicatively increasing the edge weight if it does.

Here, we initialize the edges in the border graph to weight 50, while new edges made during iteration to non-geographic neighbors are initialized with weight 1, reflecting the comparative difficulty of migration at distance. We use a "jump" probability $p = 0.05$ and select destination nodes from the entire graph excepting the source node, weighted by $\log GDP + \log population$. We decided on using GDP and population because they are among the most easily accessible data on most countries, they're intuitive indicators of world influence (and migration desirability), and they're unary traits (they don't depend on some bilateral relationship like cultural similarity or geographic distance), and we decided on log-weighting these two factors after experimentation with different log and linear combinations, as log-weighting provided good variance between countries without being too skewed. When selecting destination nodes from geographic neighbors (case $1 - p$), we sample uniformly, as weighting by GDP and population in this case gave far too much importance to overall country "size" or "influence" and outweighed geography. Finally, when updating edge weights, we multiply the existing weight by 1.2. This lets geographically close nodes preserve their "head start" of initial weight 50, which would effectively disappear over time if we used a linear update.

4 Results and discussion

Based on these parameters for 2000 iterations, we have a model that accurately models the change in the global migration network, though not necessarily the base features. We can see immediately in Fig. **??** the dramatic difference in variance in the degree distributions between the GMD and the geo-community model, though the means are roughly the same. This makes a certain amount of sense

Algorithm 3 Geocommunity model

```
procedure GEOCOMMUNITY(iters, p)
    g ← directed graph of n country nodes
    for (u, v) with u ∈ g.nodes, v ∈ g.nodes do
        if u, v share a land border then
            add edge (u, v) to g
        end if
    end for
    for i = 0; i < iters; i++ do
        for u ∈ g.nodes do
            r = uniform random value ∈ [0, 1)
            if r < p then
                v ← n ∈ weighted g.nodes
            else
                v ← node ∈ neighbors of u
            end if
            add or update edge (u, v)
        end for
    end for
end procedure
```

— in our model, the most common links (by a large amount) are made to geographic neighbors, and the range of possible number of bordering countries is very low (Russia and China tie for the most neighboring countries, at 16).

The power-law relationship is fairly well modeled here, as are the changes in density (0.322 to 0.472, vs. 0.310 to 0.466 in the GMD) and average shortest path (1.467 to 1.286, vs. 1.539 to 1.379 in the GMD). Assortativity is negative but non-negligibly higher than in the GMD (-0.007 to -0.01 from 1500 to 2500 iterations, vs. -0.205 to -0.209 in the GMD from 1960 to 2000), suggesting that the power-law relationship might be due to preferential-attachment effects — which would make sense, given the way we modeled weighted random jumps. Clustering coefficient gets close to the GMD, at 0.715, at 2500 iterations but is much lower, at 0.536, at 1500 iterations, which is interesting given the initial constraint and overall emphasis on immediate geographic neighbors; when we performed spectral clustering as in Fig. 2, the clusters appear much less geographically based.

We hypothesized several possible reasons for these discrepancies. The focus on generating and updating outlinks assumes a primarily emigration-driven evolution of the network — that is, that the primary function of nodes is to act as sources, not sinks — which might miss some set of factors influencing immigration. The strong focus on land

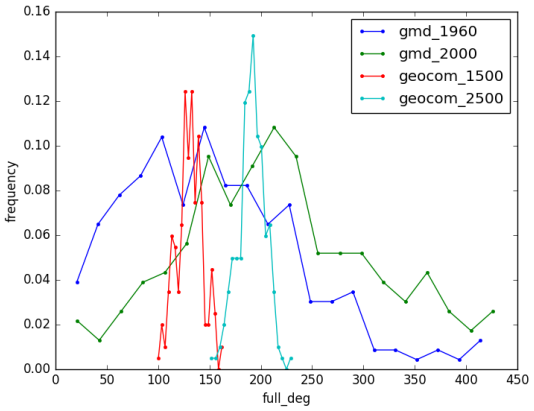

Figure 9: Degree distribution for the geocommunity model at 1500 and 2500 iterations and the GMD from 1960 and 2000.

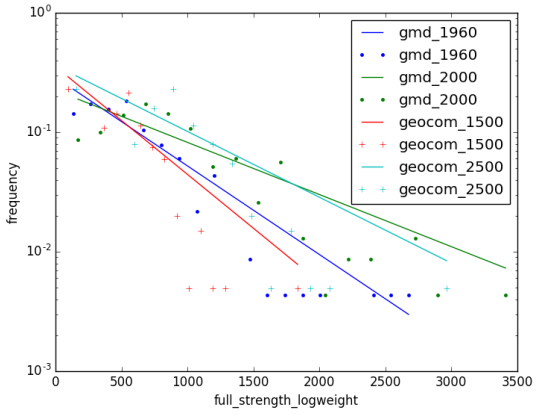

Figure 10: Node strength distribution for the geocommunity model at 1500 and 2500 iterations and the GMD from 1960 and 2000.

borders also plays a role: island nations, including many small countries but also major players like Australia, must depend entirely on the small proportion of random jumps for in-degree weight, and countries like the U.S. and the U.K., which border two countries and one country respectively, are also disproportionately affected. Groupings like the Caribbean cluster see in Fig. 2 are impossible here; unfortunately, we could not find reliable large-scale sets of relevant maritime boundaries. Finally, the model doesn't account for the self-reinforcing way migration corridors establish themselves: once some critical mass of people begins moving between two countries, regardless of geographic location the chance of more people joining that mass becomes

significantly greater. We did experiment with including established "jump" links in the "neighbors" case, but this tended to create disproportionately high edge weights for random destination nodes that were simply added early on.

We considered and experimented with several other adjustments to the model during development, including manually boosting the role of some top n countries like the U.S., adding a forest fire-type action in which a node would form links with some medium-high probability to the neighbors of its geographic neighbors, and adding a higher community-level action in which multiple countries in a geographical community would create edges at once, but these all complicated the model without creating much improvement. Though our geocommunity model ultimately does not accurately represent all elements of the global migration network, it does capture to a certain extent the shrinking, densifying, increasingly connected nature of migration.

References

[1] Fagiolo, G., Mastrorillo, M. 2013. International migration network: Topology and modeling. *Physical Review E* 88, 012812. DOI:10.1103/PhysRevE.88.012812.

[2] Leskovec, J., Kleinberg, J., Faloutsos, C. 2007. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data* 1, 2. DOI:10.1145/1217299.1217301.

[3] Porat, I., Benguigui, L. 2016. Global migration topology analysis and modeling of bilateral flow network 2006-2010. *EPL* 115, 18002. DOI:10.1209/0295-5075/115/18002.

[4] Squartini, Tiziano and Garlaschelli, Diego. 2011. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics* 13, 083001. DOI:10.1088/1367-2630/13/8/083001.

[5] Squartini, Tiziano, Fagiolo, Giorgio, and Garlaschelli, Diego. 2011. Randomizing world trade. II. A weighted network analysis. *Physical Review E* 84, 046118. DOI:10.1103/PhysRevE.84.046118.