SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Creating a surrogate commuter network from Australian Bureau of Statistics census data

Received: 6 September 2018

Accepted: 21 June 2019

Published online: 16 August 2019

Kristopher M. Fair[1], Cameron Zachreson[1] & Mikhail Prokopenko[1,2]

Between the 2011 and 2016 national censuses, the Australian Bureau of Statistics changed its anonymity policy compliance system for the distribution of census data. The new method has resulted in dramatic inconsistencies when comparing low-resolution data to aggregated high-resolution data. Hence, aggregated totals do not match true totals, and the mismatch gets worse as the data resolution gets finer. Here, we address several aspects of this inconsistency with respect to the 2016 usual-residence to place-of-work travel data. We introduce a re-sampling system that rectifies many of the artifacts introduced by the new ABS protocol, ensuring a higher level of consistency across partition sizes. We offer a surrogate high-resolution 2016 commuter dataset that reduces the difference between the aggregated and true commuter totals from ~34% to only ~7%, which is on the order of the discrepancy across partition resolutions in data from earlier years.

Background & Summary

High-resolution commuter network information, as well as general information describing population distributions[1], is a major factor in the computational modeling of diffusion phenomena in various contexts: demographic[2], epidemiological[3–6], economic[7], ecological[8] and so on. However, privacy constraints on released Census data, in the presence of intricate dependencies between population and employment distributions in relatively small, highly urbanized, but spatially spread countries, such as Australia, coupled with changes in data protocols across census years, present specific challenges in reconstructing commuter (travel-to-work) networks with sufficiently high fidelity[1,9–12].

These challenges manifest in two ways. The first of these pertains to individual microdata, which is organized by household to capture information about both the individual and housing unit. While the collective microdata is a powerful resource, variations in questions asked, possible responses, and record structure often present difficulties in comparing results across years[13]. The second challenge relates to the specific methods used by agencies that gather and report census data, in protecting the anonymity of individuals. While it is necessary for these methods to introduce perturbations, the details of how such perturbations are applied can result in unintended consequences when high-resolution data is aggregated. This is because biases introduced by the perturbation protocol are magnified by aggregation.

In the recent Australian census datasets[14], these challenges manifest themselves as loss of accuracy in very finely partitioned data, where individual population counts can be on the order 1 to 10 individuals. An important example of such a data set is the commuter network, describing the normal work travel behaviour of the population. The loss of accuracy in such data is primarily due to the specific noise-inducing protocols that the Australian Bureau of Statistics (ABS) employs to ensure the anonymity of census participants. At the same time, this loss in accuracy severely diminishes the usefulness of the commuter networks in modelling contagion phenomena, such as epidemics. In such models, work mobility is a primary driver of contagious diffusion. As such, the accuracy of the commuter network is crucial for realistic outputs regarding aggregate demographic and epidemiological characteristics, such as the community and national attack rates. Furthermore, without trustworthy inputs, such models cannot accurately identify salient routes of contagion spread, or analyze mitigation strategies based on network theory.

[1]Complex Systems Research Group, School of Civil Engineering, Faculty of Engineering, The University of Sydney, Sydney, NSW, 2006, Australia. [2]Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Westmead, NSW, 2145, Australia. Correspondence and requests for materials should be addressed to K.M.F. (email: kristopher.fair@sydney.edu.au)

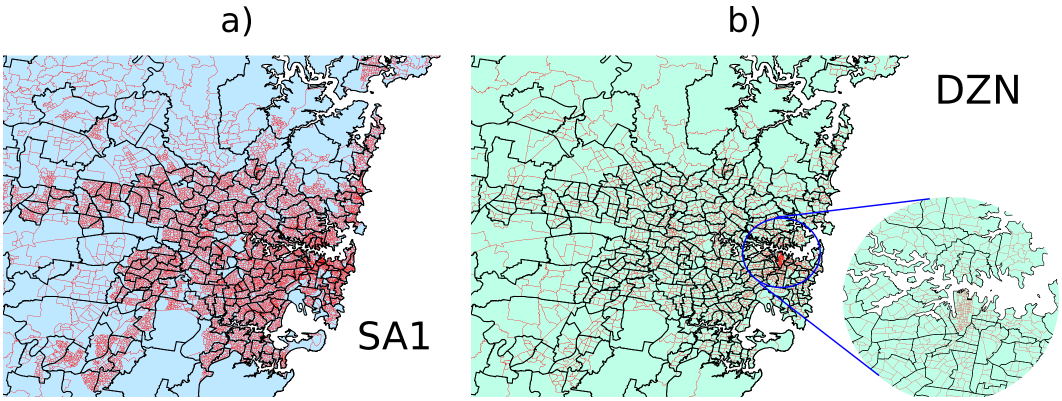

Fig. 1 Maps of the Greater Sydney region illustrating the distribution of population partitions. (**a**) A map of the Greater Sydney region showing SA2 (black) and SA1 (red) population partitions. (**b**) A map of the same area showing SA2 (black) and DZN (red) partitions. The inset in (**b**) zooms in on the Sydney central business district to illustrate the much denser packing of DZN partitions in that area.

Similar challenges from noise-inducing protocols, which may also differ across census years, occur in other scenarios in which there is a need to estimate demographic and phenomenological dynamics. This is relevant not only to network-centric studies, but also to more general agent-based simulations, or any study aimed at the fine-grained reconstruction of spatio-temporal dynamics[15]. Thus, the goal of the present work is not only to reconstruct the specific commuter networks of Australia between 2011 and 2016, but also to present a method of microdata reconstruction. The method aims to correct discrepancies that may arise due to the noise protocols used to anonymize the Census, improving consistency across partition scales while preserving anonymity. The secondary aim is to increase the interoperability of Census datasets, in line with the Integrated Public Use Microdata Series (IPUMS) approach[13].

To further these ambitions we first formalize the network structure and identify discrepancies between different scales of spatial partitioning. We then describe the technical details for constructing our re-sampled network using additional datasets. Finally, we show several comparisons between the ABS provided and the re-sampled data that demonstrate the distinction and validity of the resulting dataset.

The ABS provides access to most census data through the on-line system Census TableBuilder, free of charge, for the 2006 census onward. A subset of the available data is the accumulated microdata of usual-residence (UR) to place-of-work (POW) which constitutes the commuter mobility network (we will refer to this as the TTW, or, travel-to-work dataset). Each census has undergone some re-partitioning of residential and work areas with the latest hierarchical structure divided into four levels of statistical areas for UR (UR = [SA1, SA2, SA3, SA4]), and POW (POW = [DZN, SA2, SA3, SA4]), respectively. This system is defined by the Australian Statistical Geography Standard[16]. The smallest of these residential partitions, SA1, is designed to house a population of about 200 to 800 people. Maps of SA2, SA1, and DZN partitions for the Greater Sydney region are displayed in Fig. 1. SA1 and DZN partitions accumulate to exact partitions on the SA2 scale, this is displayed for SA1 partitions in Fig. 1a, and for DZN partitions in Fig. 1b. This exact correspondence allows unambiguous amalgamation of statistics from smaller to larger spatial scales. Note that the uneven distribution of employment centers in Australia's cities produces a corresponding non-uniformity in DZN partition density, as displayed in Fig. 1b.

This partitioned commuter data translates to a bipartite network $G_{[\mathrm{UR}\rightarrow\mathrm{POW}]}=(V_G, E_G)$ where V_G is a set of vertices (nodes) of two types $V_G = X \cup Y$, where $X = \{x_1, x_2, \ldots, x_n\}$ represent the n partitioned UR locations, and $Y = \{y_1, y_2, \ldots, y_k\}$ represent the k partitioned POW locations. The set of edges

$$E_G = \{(x_{i_1}, y_{j_1}),\ (x_{i_2}, y_{j_2}),\ \ldots(x_{i_{|E_G|}},\ y_{j_{|E_G|}})\}, \tag{1}$$

defines the unique connections between these vertices. For example, UR x_i and POW y_j may be connected by an edge $e_{ij} = (x_i, y_j)$. Each subset of edges has a corresponding set of weights, defined by the function:

$$w_{ij}(\{e_{ij}\},\ G), \tag{2}$$

which gives a set of commuter numbers indexed to the corresponding location pairs in $\{e_{ij}\}$, over the network G. The use of the argument G is necessary, as the same location pairs may have different numbers of commuters in different networks. For brevity, we will omit the subscripts i and j in cases where they are not required for specificity. We will use similar notation to refer to the sets of UR and POW locations associated with edge sets $\{e\}$ as $x(\{e\})$ and $y(\{e\})$, as well [Note: the second argument is not necessary here, as the required information is contained in the set $\{e\}$, and does not vary between networks with the same sets of nodes].

As mentioned above, these data sets are subject to a perturbation protocol to prevent cross referencing different variables that may allow the identification of specific individuals[17] even with the application of safeguards[18,19]. Not doing so would violate the Australian Census and Statistics Act 1905 to preserve the anonymity of individuals. The ABS applies two general categories of perturbations, *data suppression* in which tabular data that presents a high risk of cross-identification is simply removed, and *data modification*. Data suppression typically involves

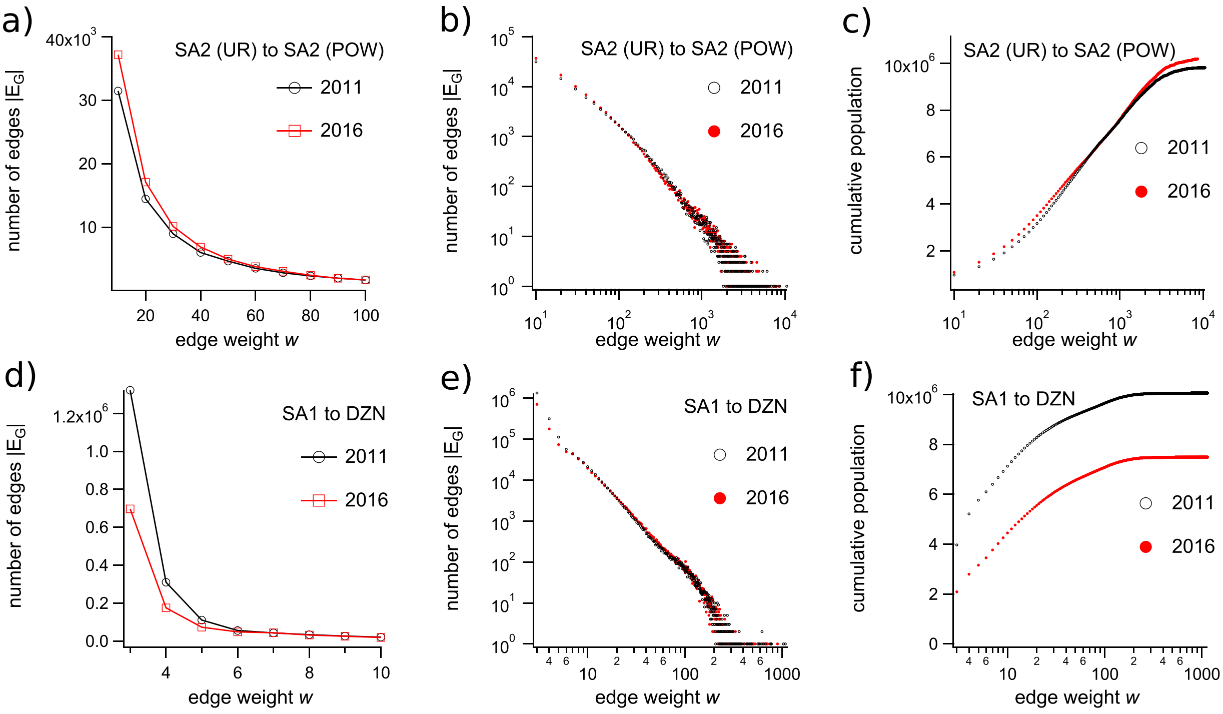

Fig. 2 Weight distributions and cumulative population distributions for TTW networks from different census years and partition schemes. (**a**) Distributions of edge weights ($w < 100$) for the SA2 → SA2 networks for 2011 and 2016, plotted on a linear scale. (**b**) Distributions of all edge weights for the SA2 → SA2 network from 2011 and 2016 plotted on a log scale. (**c**) Cumulative population distributions for the SA2 → SA2 network from 2011 and 2016. (**d**) Distributions of edge weights ($w < 10$) for the SA1 → DZN networks for 2011 and 2016, plotted on a linear scale. (**e**) Distributions of all edge weights for the SA1 → DZN network from 2011 and 2016 plotted on a log scale. (**f**) Cumulative population distributions for the SA1 → DZN network from 2011 and 2016. The distributions in (**a**–**c**) have bin width of 10, while (**c**,**d**) have bin width 1, with a minimum value of 3, artificially introduced by the ABS protocol. The plots in (**a**,**d**) show only a subset of the weight range, zooming in on the low end of the distribution where the largest discrepancies exist between years.

the introduction of a high-pass threshold below which entries are set to zero. Data modification involves various methods of additive noise perturbation as discussed in ABS publications[20–22]. [Note: The referenced ABS publications contain various descriptions of perturbation methods that may or may not have actually been applied by the ABS whose data perturbation policies are subject to change and may vary between data sets.]

The sizes of UR and POW population partitions affect the magnitudes of the populations moving between them. Relative to these magnitudes, different levels of noise addition and data suppression are required to preserve the anonymity of individuals. Furthermore, for the 2016 census, the ABS changed their perturbation protocol by removing a step designed to conserve the total population across different spatial partitions, a property they refer to as 'additivity'. Some major practical consequences of removing the additivity-ensuring step are observable discrepancies in the total number of commuters, $N_G = \sum w(E_G, G)$, accounted for by the network G on different partition scales.

Edge weight distributions, and cumulative population distributions as functions of edge weight for the SA2 → SA2 and SA1 → DZN commuter networks of 2011 and 2016 are displayed in Fig. 2. Lower-resolution TTW networks such as those representing connections on the SA2 scale display relatively consistent weight distributions between censuses. Comparison across years shows moderate increases in the numbers of edges across the weight range as could be expected for an increasing employed population between 2011 and 2016 (Fig. 2a,b). The corresponding distribution of this increased population across the edge weight range is illustrated in Fig. 2c, which does not show any alarming trends or obvious artifacts in the data. Unfortunately, this consistency does not hold for the fine-grained SA1 → DZN network. The weight distributions for this network shown in Fig. 2d,e indicate a counter-intuitive drop in the numbers of small edges between 2011 and 2016, which corresponds to a dramatic decrease in the total commuting population accounted for by the network. The distribution of the commuting population across the edge weight range (Fig. 2f) confirms that major discrepancies exist between partition schemes, likely due to a significant drop in the number of small edges included in the network.

As the partitions that comprise the vertices V_G are increasingly subdivided, the weights of the edges connecting them get smaller. The new perturbation protocol appears to dramatically reduce the number of small edges included in the network, particularly around the minimum value of $w = 3$. This adversely effects the network both quantitatively, by lowering the commuter populations throughout the network, and structurally, by removing edges from E_G, which alters the binary structure of the network. In the case of the high-resolution SA1 → DZN

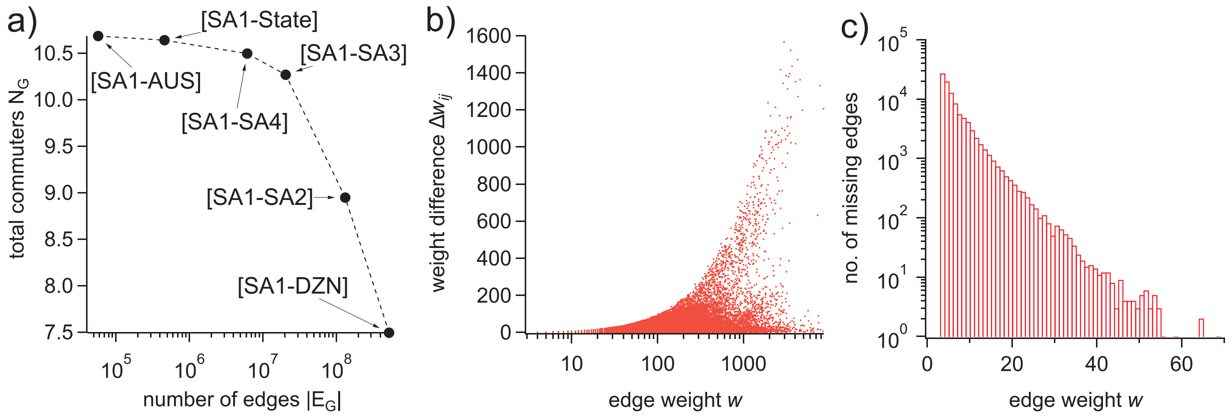

Fig. 3 Discrepancies in the total population and commuter distribution related to partition aggregation behavior. (**a**) The total number of commuters N_G in ABS data for networks of varied size. Each point corresponds to a network between SA1 partitions and a different scale of POW partition (national, state, SA4, SA3, SA2, DZN). (**b**) The discrepancy between commuter numbers, Δw_{ij}, on each edge $w(E_{AB}, A)$ and $w(E_{AB}, B)$ plotted against $w(E_{AB}, B)$. (**c**) The frequency distribution as a function of edge weight for edges present in the ABS-provided SA2 → SA2 network (B) but not the aggregated SA1 → DZN network (A).

network, small edges are a crucial aspect of the network structure, and carry a large portion of the total edge strength.

The need for a method to ensure consistency in commuter numbers across partition scales is further exemplified in Fig. 3a, which plots the total working population (N_G) in networks built by distributing commuters from SA1 partitions into each of the possible POW partition schemes. As the sizes of the POW partitions decrease from the entire nation down to individual destination zones, the total number of commuters drops by 34% while the total number of edges increases by four orders of magnitude.

The structural inconsistency across partition scales that this problem introduces can be understood by amalgamating the vertices of network $G_{[\mathrm{SA1} \to \mathrm{DZN}]}$ into the corresponding SA2 partitions. By doing so, we create network $A_{[\mathrm{SA2} \to \mathrm{SA2}]} = (V_A, E_A)$, that can be compared to the network constructed from ABS data on the SA2 scale [which we will label network $B_{[\mathrm{SA2} \to \mathrm{SA2}]} = (V_B, E_B)$]. Network B is missing only 6% of the total commuter population because the edges are composed of more commuters and therefore receive relatively less perturbation from the ABS protocol. This smaller discrepancy is comparable with that of previous years for which the additivity-ensuring step was still included.

Figure 3b illustrates the discrepancies between edge weights (commuter numbers between a given pair of locations) for edges appearing in both networks A and B. To compute these discrepancies, we define the set of edges appearing in both E_A and E_B as the intersection $E_{AB} = E_B \cap E_A$, the weights of these edges for networks B and A, respectively, as $\mathbf{w}_B = w(E_{AB}, B)$, and $\mathbf{w}_A = w(E_{AB}, A)$, and the discrepancies Δw between the weights of edges existing in both sets

$$\Delta w_{ij} = [w_{ij} \in \mathbf{w}_B] - [w_{ij} \in \mathbf{w}_A]. \tag{3}$$

Using this notation, Fig. 3b plots Δw_{ij} as a function of $w \in \mathbf{w}_B$, and demonstrates that the perturbations to small edges in the SA1 → DZN network produce large negative discrepancies in edge weight when the data is aggregated to the SA2 → SA2 scale.

To understand this result in more detail, it is helpful to note that the spatial distribution of the working population is very heterogeneous, with an exponentially larger fraction of the working population employed within the central business districts of major cities. However, only the DZN partitions are designed to accommodate this heterogeneity, as they are delineated based on employee population (the number of people who work in a region), rather than residential population. On the other hand, SA2 partitions are designed based on residential population which results in a few SA2 business hubs containing many DZN partitions (see Fig. 1b). In some cases, over 10^3 component SA1 → DZN edges amalgamate to single, larger SA2 → SA2 edges.

It is clear that many SA1 → DZN edges are being removed entirely (their weight set to zero) because there are 97,881 edges appearing in the as-provided SA2 → SA2 network B that do not appear after aggregating the SA1 → DZN edges to produce network A. This gives $|E_A| \approx 0.64|E_B|$ for the SA2-level networks. The frequency distribution for the weights of missing edges, $w(\{E_B \backslash E_A\}, B)$ (where the symbol\denotes the set complement), is shown in Fig. 3c which indicates an exponential decrease in removal frequency as a function of edge weight. The data in Figs 2 and 3 indicate conclusively that many small perturbations on the SA1 → DZN scale accumulate, producing the large discrepancies observed when they are aggregated.

In this work, we develop and apply a method to restore the lost network structure and improve quantitative consistency across commuter networks on different partition scales. The result is a surrogate network $S_{[\mathrm{SA1} \to \mathrm{DZN}]} = (V_S, E_S)$, on the resolution of SA1 to DZN. This reconstructed commuter network will serve as a platform for ongoing research efforts that utilize Australian travel networks, such as agent-based epidemiological modeling[5,23].

Network	Partition (UR → POW)	$\|E\|$	$\sum w$	Source
$R=(V_R, E_R)$	SA1 → DZN	1,184,946	7,023,571	ABS 2016
$A=(V_A, E_A)$	SA2 → SA2	118,167	7,023,571	Accumulated from R
$B=(V_B, E_B)$	SA2 → SA2	212,805	10,073,246	ABS 2016
$\Gamma=(V_\Gamma, E_\Gamma)$	SA2 → DZN	515,250	9,853,543	ABS 2016
$H=(V_H, E_H)$	SA1 → DZN	2,046,094	10,058,331	ABS 2011
$S=(V_S, E_S)$	SA1 → DZN	1,731,938	9,336,333	Constructed
$C=(V_C, E_C)$	SA2 → SA2	118,167	9,336,333	Accumulated from S

Table 1. Commuter networks and selected characteristics.

Methods

Our method is essentially a re-sampling process that we use to introduce new edges into the SA1 → DZN network to improve quantitative consistency upon aggregation to the SA2 scale. The method does not introduce any new edges to the SA2 → SA2 network upon aggregation, and therefore cannot correct for the missing edges distributed as shown in Fig. 3c. However, most of the missing commuters are accounted for by correcting the discrepancies shown in Fig. 3b, and our method emphasizes this aspect of the problem.

Before commencing our procedure, we pre-processed all data provided by the ABS to remove the edges that link to non-geographic regions such as "Migratory/offshore/shipping" and "No usual address". For the 2016 SA1 → DZN network this accounts for 53,135 edges and 469,854 commuters.

In addition to the original, perturbed SA1 → DZN network, the method requires the following sets and quantities that we obtained from independent ABS databases:

- $N_X = \left\{N_{x_1},\ N_{x_2},\ \ldots N_{x_n}\right\}$ and $N_Y = \left\{N_{y_1},\ N_{y_2},\ \ldots N_{y_n}\right\}$, the set of local worker populations for SA1 and DZN partitions, respectively.
- The SA2 → SA2 commuter numbers from the ABS-provided SA2 → SA2 network (B).
- The set of (unweighted) SA2 → DZN edges found by creating a mixed-partition network.
- $P(w|N_x)$, the normalized distribution of edge weights w given residential population N_x.

The last item refers to the relationship between the local distribution of edge weights and the population of the associated SA1, as calculated from 2011 census data obtained without the updated privacy policy compliance protocol.

Our method can be summarized as a two-step process:

1. Produce a set of q candidate out-edges $M = \{m_1,\ m_2,\ \ldots m_q\} = \left\{(x_{i_1},\ w_1),\ (x_{i_2},\ w_2)\ldots\left(x_{i_q},\ w_q\right)\right\}$, specifying the SA1 ($x$) and the number of commuters (w). This set accounts for the missing workers from each SA1 while maintaining a realistic dependence of the weight distribution on the UR population $P(w|N_x)$.
2. Build network S: add the candidate edges in M into the SA1 → DZN network by specifying a DZN (y) without violating the topology of the SA2 → DZN network, exceeding the population of the DZN, adding edges that are not present in the SA2 → SA2 network, or exceeding the known commuter populations between locations in the SA2 → SA2 network.

In addition to the networks A, B, and S defined above, we will refer to several distinct network sets that are important for the explicit description of our process. For clarity, we will summarize these here and give a brief description of their role in our method.

Network R is the ABS-provided SA1 → DZN network (referred to above as $G_{[SA1\rightarrow DZN]}$), which was released by the ABS subject to the perturbations this work is intended to correct. Network A is the SA2 → SA2 network aggregated from R. Network B is the ABS-provided SA2 → SA2 network that exhibits relatively consistent aggregation behavior (that is, the total number of commuters it accounts for is roughly 94% of the known total). We use network B as a quantitative ground-truth while generating the surrogate network. Network H is the ABS-provided SA1 → DZN network from the 2011 census, which exhibits acceptable aggregation behavior. We use network H to build up the set of probability distributions describing $P(w|N_x)$. A key assumption of our method is that this relationship between local population and out-edge weight distribution is relatively invariant across census years. Network Γ is the ABS-provided SA2 → DZN network which we use as a topological constraint while assigning the candidate edges from each residential zone to appropriate destination zones. That is, we only incorporate SA1 → DZN edges into S that have a corresponding SA2 → DZN pair existing in Γ. Finally, network S is the surrogate SA1 → DZN network that is the final output of our method and network C is the SA2 → SA2 network aggregated from network S. We compare networks B and C when evaluating the aggregation behaviour of S. Some quantitative features of these networks are summarized in Table 1.

The following two sections describe our method in detail. The first describes the process of generating the list of (SA1, w) pairs which we refer to as the "candidate edges". The second describes the process of assigning these candidate edges to DZN partitions subject to our selected constraints.

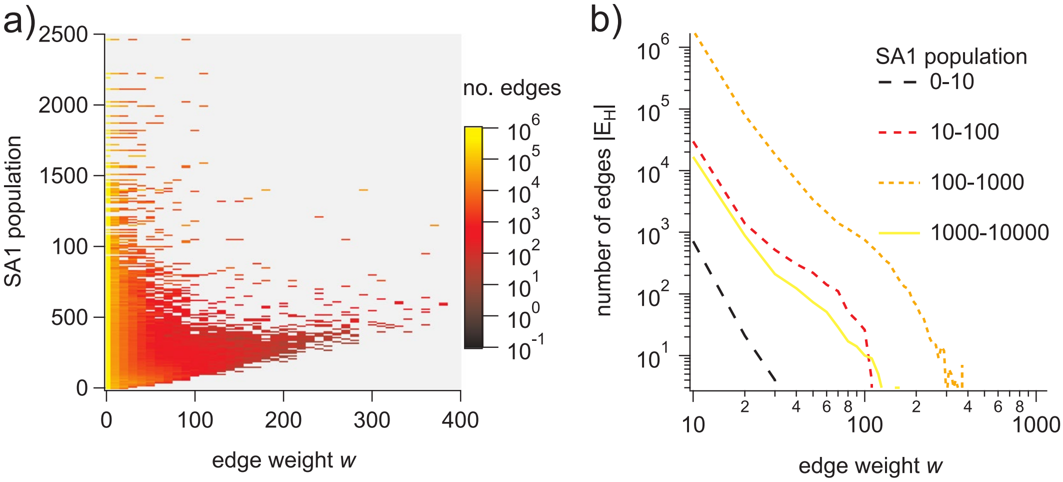

Fig. 4 Edge weight frequency distributions as functions of the local population. (**a**) Color plot showing $P(w)$ (y axis) as a function of N_x (x axis) for the 2011 SA1 → DZN commuter network. (**b**) The frequency distribution of edges as a function of the SA1 → DZN commuter network edge weight, where each curve represents the weight frequency distribution for a specific range of SA1 populations.

SA1 candidate edges. We observed the behavior of $P(w)$ as a function of N_x to be similar across 2006 and 2011 censuses. This dependence appears to reflect a consistent feature of the commuter mobility network. Although the underlying mechanism producing this set of conditional distributions is not in the scope of this report, it is a subtle aspect of the network structure that should be taken into account. Network $H_{[\mathrm{SA1}\rightarrow\mathrm{DZN}]} = (V_H, E_H)$, derived directly from the 2011 ABS census, along with the 2011 worker populations, gives the distribution of commuter edge weights as a function of the local SA1 population $P(w|N_x)$ (shown in Fig. 4). While the method we used to generate these distributions is case-specific, a similar process could be applied in any situation where there is some confidence in the separation of time-scales between real network evolution and artifact introduction due to institutional data processing protocols. Indeed, a more general approach to this aspect of the problem may be needed in cases where true network dynamics are more difficult to distinguish from artifacts. This is an ongoing question that we will continue to address in future work. One promising future direction is to derive a maximum entropy distribution for the weights of the edges leaving each location, constrained by the known numbers of commuters and the worker populations in the destination zones allowed by the topology and SA2 → DZN edge weights of network Γ. In general, the maximum entropy principle determines the least biased probability distributions, consistent with specific constraints on the average values of measurable quantities[24]. Other approaches are possible as well, for example, Shannon information could be computed for fragments of the network that exhibit acceptable aggregation behavior, and local weight distributions defined so that sampling from them explicitly addresses information loss in parts of the network adversely affected by the removal of data from the original travel-to-work matrix. Techniques for doing so could be adapted from existing methods where networks are iteratively grown from fragments based on node assortativity constraints, leveraging the relationships between node assortativity and mutual information of the target network[25,26].

Once these conditional distributions are established, we sample from them to account for the number of missing commuters from each SA1. The number of missing commuters associated with a given SA1 partition x^* is computed as the discrepancy between the known working population (N_{x_i}) and the sum $\sum_{j=1}^{k} w\left(\left\{\left(x^*, y_j\right)\right\}, R\right)$, which is the total out-weight associated with the partition x^*. The set of these accumulated populations gives N_{X_R}:

$$N_{X_R} = \left[\sum_{j=1}^{k} w\left(\left\{\left(x_1, y_j\right)\right\}, R\right), \ldots \sum_{j=1}^{k} w\left(\left\{\left(x_n, y_j\right)\right\}, R\right)\right] = \left\{N_{x_1}^R, N_{x_2}^R, \ldots N_{x_n}^R\right\}, \tag{4}$$

which allows us to calculate the discrepancy in the local worker population for each SA1:

$$\Delta N_X = \left\{\left[N_{x_1} - N_{x_1}^R\right], \left[N_{x_2} - N_{x_2}^R\right], \ldots \left[N_{x_n} - N_{x_n}^R\right]\right\} = \left\{\Delta N_{x_1}, \Delta N_{x_2}, \ldots \Delta N_{x_n}\right\}. \tag{5}$$

The algorithm then generates M as follows: for each SA1 partition x_i, individual weights w' are iteratively sampled from $P\left(w\middle|N_{x_i}\right)$ to produce the candidate edges $m' = (x_i, w')$ which are included in M under the condition that

$$\Delta N_{x_i} > w' + \sum_{m_j \in M} w_j \times \delta\left(x_{i_j}, x_i\right), \tag{6}$$

where $\delta\left(x_{i_j}, x_i\right)$ is equal to 1 if $x_{i_j} = x_i$ and equal to 0 otherwise. If the condition above is not met the candidate edge m' is rejected. The sampling process is repeated until the discrepancies ΔN_{X_n} are all less than three, the

Set	Contents	Set size	Total population	Source
$M=\{m_1, m_2, \dots m_q\} = \left\{(x_{i_1},\ w_1),\ (x_{i_2},\ w_2),\ \dots\left(x_{i_q},\ w_q\right)\right\}$	SA1 candidate edges	683,239	2,572,117	Constructed
$N_X = \left\{N_{x_1},\ N_{x_2},\ \dots N_{x_n}\right\}$	SA1 employed residents	57,523	10,113,273	ABS 2016
$N_Y = \left\{N_{y_1},\ N_{y_2},\ \dots N_{y_k}\right\}$	DZN employees	9,151	10,677,111	ABS 2016

Table 2. Independent data sets and selected characteristics.

smallest edge size. That is, candidate edges are generated to precisely account for the number of workers missing from each SA1. Quantitative features for an instance of the candidate edge set M, and the local populations used to constrain its construction (N_X) and assignment (N_Y) are shown in Table 2. The algorithmic process for creating the set of candidate edges is outlined by the pseudocode in Box 1. The following section describes the process of assigning the candidate edges to destination zones.

Assigning edges. Once the set of candidate edges is generated, each specifying an edge weight and SA1 origin vertex, all that remains is to assign them DZN vertices. Then, the new edges can be included in the network R to create the surrogate network S. The procedure we used for these assignments is described in this section and outlined in Box 2.

We assign candidate edges from M to reasonable DZN partitions by employing $\Gamma_{[\mathrm{SA2}\rightarrow\mathrm{DZN}]}$, $B_{[\mathrm{SA2}\rightarrow\mathrm{SA2}]}$, E_{AB}, and N_Y to conditionally restrict the connections that can be added in order to maintain the lower-resolution topology and the worker populations at the destination zones. The networks Γ and E_{AB} are used as binary topological constraints, restricting the possible set of {SA2, DZN} and {SA2, SA2} location pairs that are compatible with the topology of the new network E_S. We use Γ as a topological constraint because it represents a good compromise between resolution and quantitative consistency. Because of the larger partitioning of the residential zones X_Γ, the network loses approximately 8% of total commuters due to ABS perturbations, which is much better aggregation behavior than we observe on the SA1 → DZN scale, but worse than the SA2-level network on these terms. On the other hand, it explicitly accounts for the connectivity between SA2 residential partitions and DZNs, making it a stronger constraint than the SA2 → SA2 network. We use the overlapping edge set E_{AB} as a topological constraint because it restricts our procedure to those parts of the network in which we have the most confidence. We take this conservative approach in order to avoid introducing edges to the network that could artificially increase connectivity across disparate regions. The local worker populations at each DZN (N_Y) are used as quantitative constraints, ensuring that the local populations are not exceeded due to the addition of new edges. Similarly, $w(E_{AB}, B)$, the number of commuters between SA2(UR) and SA2(POW) in the portions of network B that overlap with A, constrains the number of commuters that can be added to particular edges in S.

To select SA1 vertices for the candidate edges M, we iterate through the DZN partitions and perform the following procedure:

For each DZN destination vertex y_i we use Γ and E_{AB} to determine the set of possible SA1 origin vertices. These define the subset $M' \subseteq M$ compatible with both the SA2 → DZN and SA2 → SA2 topologies. We then sample M' uniformly at random, combining the sample with the current destination zone y_i to produce a new edge. The new edge is added to the surrogate network under the condition that doing so does not exceed the known number of commuters between SA2 partitions when the surrogate network is aggregated.

To be precise, Γ, E_{AB}, and y_i define the subset of SA2 → DZN edges

$$E'_\Gamma = \left\{e \in E_\Gamma \middle| y(\{e\}) = y_i,\ \left(x(\{e\}),\ \Upsilon_{y_i}\right) \in E_{AB}\right\}, \tag{7}$$

where Υ_{y_i} is the SA2 partition containing the DZN y_i. In words, E'_Γ is the set of SA2 → DZN edges that point to the destination zone y_i and are consistent with the SA2 → SA2 topology E_{AB}. These define the SA2 partitions $\Phi_i = x(E'_\Gamma)$ and the subset of SA1 partitions contained by them which we will call X_{Φ_i}. From these, the subset of candidate edges is simply determined by selecting only those that contain an element of X_{Φ_i} as origin vertex:

$$M' = \left\{m_j \in M \middle| x_{i_j} \in X_{\Phi_i}\right\}. \tag{8}$$

Once M' is defined, we randomly select a candidate $m^* \in M' = (x^*, w^*)$ with uniform probability, producing a potential new edge $e^* = (x^*, y_i)$ with weight $w(e^*) = w^*$. The new SA1 → DZN edge e^* aggregates into the SA2 → SA2 edge

$$e_B = \left\{e \in E_B \middle| X_x \supseteq x(\{e^*\}),\ Y_y \supseteq y_i\right\} = (x_B,\ y_B), \tag{9}$$

where X_x and Y_y are the sets of SA1 and DZN zones contained (respectively) by the SA2(UR) and SA2(POW) partitions in each element of E_B.

To check whether or not the new edge e^* should be added to the surrogate network, we aggregate E_S over the SA1 and DZN vertices contained by the SA2 partitions x_B and y_B, and determine whether adding the new edge will exceed the known number of commuters between the SA2 zones. That is, the edge e^* is added to E_S under the condition that

Box 1. The candidate edge set algorithm. Pseudocode for the algorithm that produces a list of candidate edges from each SA1 that match the local commuter populations and dependence of edge weight distribution on the worker population as determined by the 2011 census.

```
procedure GENERATE CANDIDATE EDGES
input:
$N_{X_R}$, the number of SA1 employees aggregated from $R$
$N_X$, the number of SA1 employees reported by ABS
$P(w|N_x)$, the 2011 edge weight distribution conditional on local population
for
 $x_i$ in $X_R$:
$N_{x_i}^R = \sum_m^{j=1} w\left(\left\{\left(w_i, y_j\right)\right\}, R\right)$
$\varDelta N_{x_i} = N_{x_i} - N_{x_i}^R$, the number of employees remaining unassigned from $x_i$
while
 $\varDelta N_{x_i} > 3$ do:
$w' = \textit{sample w with probability } P\left(w \middle| N_{x_i}\right)$
if:
 $\Delta N_{x_i} \geq w'$
$m' = (x_i, w')$
append $m'$ to $M$
$\Delta N_{x_i} - = w'$, subtract $w'$ from $\Delta N_{x_i}$
end if
end while
end for
```

Box 2. The destination assignment algorithm. Pseudocode for the algorithm that links the candidate edges to DZNs partitions, producing the surrogate network *S*.

```
procedure ASSIGNING CANDIDATE EDGES
input:
$E_B$, the SA2(UR) → SA2(POW) network reported by ABS
$\Gamma$, the SA2 → DZN network
$M$, the candidate edges produced by Algorithm 1
$R$, the SA1 → DZN network reported by ABS
$N_Y = \left\{N_{y_1}, N_{y_2}, \ldots N_{y_k}\right\}$, the DZN employee population
initialize $S = R$
initialize $\{\Delta w\}$, the discrepancies in aggregated commuter numbers (see equation 3)
while
 $|M| > 1$
for
 $y_i$ in $Y_R$:
$E_\Gamma' = \left\{e \in E_\Gamma \middle| y(\{e\}) = y_i, \left(x(\{e\}), \Upsilon_{y_i}\right) \in E_{AB}\right\}$ (equation 7)
$\Phi_i = x(E_\Gamma')$, the SA1 partitions contained by the SA2(UR) partitions of $E_\Gamma'$
$M' = \left\{m_j \in M \middle| x_{i_j} \in X_{\Phi_i}\right\}$, subset of $M$ such that $\Phi_i$ contains $x_{i_j}$
$\textit{sample } m^* = (x^*, w^*) \textit{ from } M' \textit{ uniformly at random}$
$e^* = (x^*, y_i)$, $w(\{e^*\}) = w^*$, the potential new SA1 → DZN edge
$e_B = \{e \in E_B \,|\, X_x \supseteq x(\{e^*\}), Y_y \supseteq y_i\} = (x_B, y_B)$ (equation 9)
if:
 $w(\{e^*\}) > \Delta w(\{e_B\})$ AND $N_{y_i} \geq w(\{e^\star\}) + \sum_{p=1}^n w\left(\left\{(x_p, y_i)\right\}, S\right)$
append $e^*$ to $E_S$
$\Delta w(\{e_B\}) -= w(\{e^*\})$
end if
end for
end while
```

$$w(\{e_B\}) \geq w(\{e^\star\}) + \sum_{e_{ij} \in E_S} w\left(\{e_{ij}\}, S\right) \times \delta\left(e_{ij}, X_{x_B}, Y_{y_B}\right), \tag{10}$$

where X_{x_B} and Y_{y_B} are the sets of SA1 and DZN partitions contained by the SA2(UR) and SA2(POW) zones specified by x_B and y_B, respectively, and

$$\delta\left(e_{ij},\, X_{x_B},\, Y_{y_B}\right) = \begin{cases} 1, & \text{if } x_i \in X_{x_B} \text{ AND } y_j \in Y_{y_B} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

To summarize, our algorithm allows the addition of e^* to E_S if the aggregation of E_S to larger partitions only produces edges that already exist in E_Γ and E_{AB}, these topological constraints are illustrated in Fig. 5. The aggregated edge weights are constrained as well, so that the addition of $w(\{e^*\})$ does exceed the value given by $w(\{e_B\}, B)$ upon aggregation of E_S to the SA2 → SA2 scale. After the successful assignment of edge e^* into E_S, the candidate edge m^* is removed from M and the process is repeated until edges meeting this condition cannot be found.

In principle, the above criterion is sufficient to ensure self-consistency across differently-partitioned data sets, however, the criteria must still account for the effect of the privacy policy compliance perturbations. To account for possible mismatch between employee numbers, we added the additional criterion that the number of workers assigned to destination y_i must not exceed the local worker population $N_{y_i} \in N_Y$. Therefore, the condition

$$N_{y_i} \geq w(\{e^\star\}) + \sum_{e_{ij} \in E_S} w\left(\{e_{ij}\},\, S\right) \times \delta\left(y\left(\{e_{ij}\}\right),\, y_i\right), \tag{12}$$

must be met, or the edge is not added to E_S. Here, $\delta(y(\{e_{ij}\}), y_i)$ is equal to 1 if $y(\{e_{ij}\}) = y_i$, and equals 0 otherwise.

Of the 2,572,117 commuters accounted for by the full set of 683,239 candidate edges M, there were 729,209 commuters comprising 61,855 edges remaining unassigned when our process terminated due to an inability to assign edges under the above criteria. Two factors are responsible for the inability of the algorithm to assign these edges. The first is that the privacy protocol, by design, ensures cross referencing totals do not match in perturbed data released by the ABS. The second is that our ground-truth topology omits the non-overlapping set $w(\{E_B \backslash E_A\}, B)$, therefore, the 612,215 missing commuters tabulated in Fig. 3c cannot be accounted for by our re-sampling procedure.

This surrogate network (S) has an additional 546,992 SA1 → DZN edges, a 25% increase as compared to network R, with a total number of commuters $N(S)$ comparable to that of the SA2 → SA2 network, $N(B)$. The total number of commuters in the as-provided SA1 → DZN network $N(G)$ is 7,023,571 the total for the surrogate network $N(S)$ is 9,336,333 and our quantitative ground-truth $N(B)$ is 10,073,246.

Data Records

We have made an instance of the reconstructed surrogate commuter network publicly available[27]. All of the data sets we used, including the original SA1 → DZN commuter mobility network, the SA2 → DZN network, the SA2 → SA2 mobility network, the number of employees in each SA1 (N_X), the number of employees in each DZN (N_Y), the SA1 to SA2 correspondence files, and the DZN to SA2 correspondence files are publicly available for both 2011 and 2016 through either Census TableBuilder (http://www.abs.gov.au/websitedbs/D3310114.nsf/Home/2016%20TableBuilder) or the ABS website (http://www.abs.gov.au/). The 2011 SA1 → DZN network (H) is no longer publicly available with the additivity-including privacy policy compliance protocol so we provide the version we used along with our surrogate network. The stability of the files available through ABS may vary with time, as evident in the removal of the additivity-ensuring step from the perturbation protocol used for all presently distributed data. To ensure reproducibility, all necessary input data sets, which were subject to our pre-processing procedure to remove non-geographic partitions, are available in our script input file located on the Zenodo repository[27] (see Usage Notes below).

Technical Validation

To quantitatively assess the aggregation behavior of the surrogate network S, we first accumulated its component edges into the corresponding SA2 → SA2 topology (which we will refer to as network C). This new aggregated surrogate network was then compared to both the ABS-provided SA2 → SA2 network and the aggregate of the original SA1 → DZN network (A), by several different metrics. To assess the overall agreement between the three networks, we first translated their edge lists and weights into adjacency matrices (Fig. 6a), and computed the 2D correlation coefficient between each pair:

$$r(\alpha,\, \beta) = \frac{\Sigma_m \Sigma_n (\alpha_{mn} - \bar{\alpha})(\beta_{mn} - \bar{\beta})}{\sqrt{3\Sigma_m \Sigma_n (\alpha_{mn} - \bar{\alpha})^2 \Sigma_m \Sigma_n (\beta_{mn} - \bar{\beta})^2}}, \tag{13}$$

where α and β represent each of the two adjacency matrices being compared. This comparison demonstrates a high degree of similarity between all three networks, with a significant improvement in correlation between the ABS-provided SA2 → SA2 network and the accumulated surrogate (Table 3). [Note: For technical validation purposes, we treat the SA2 → SA2 network as a unipartite weighted graph, even though the TTW matrix it represents is in actuality a bipartite network. The results here should be interpreted as a quantitative comparison only, as they do not analyze the bipartite structure and therefore do not represent the functional properties of the network].

Plotting the frequency distribution of edge weights for the ABS-provided SA1 → DZN commuter networks of 2016 and 2011, along with the corresponding distribution for the surrogate network (Fig. 6b) indicates a partial repair of the discrepancy in low-weight ($w < 10$) edge numbers observed between 2011 and 2016 (Fig. 2d).

The discrepancies in edge weights between the amalgamated surrogate network (C) and the ABS-provided SA2 → SA2 network (B) are plotted in Fig. 6c as functions of the edge weight from network B. Comparison of these discrepancies to those plotted in Fig. 3b indicates a dramatic improvement, comparable to the corresponding discrepancies computed for the 2011 commuter network. To further demonstrate the structural repair imparted to the surrogate network, we computed the distributions of the weighted degrees (the sum of all edge

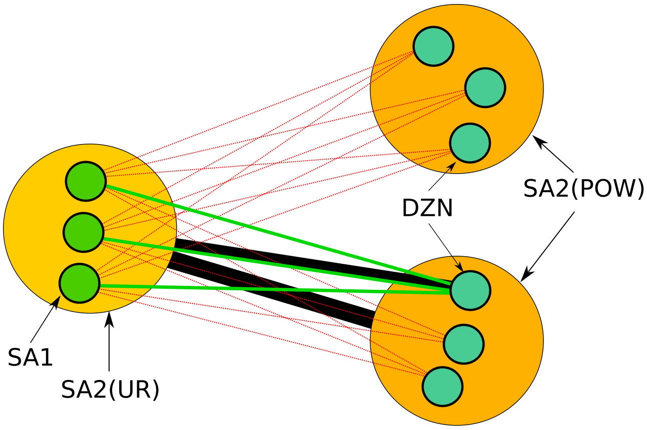

Fig. 5 Schematic of the topological constraints applied when adding new edges to the surrogate network. The black lines represent the known SA2 → SA2 and SA2 → DZN connections given by the networks B and Γ. The green lines are the allowed surrogate SA1 → DZN edges, as they are consistent with the known larger-scale topology. The red lines represent edges that are not allowed, as their inclusion would violate our constraints after aggregation of the surrogate to larger partition schemes.

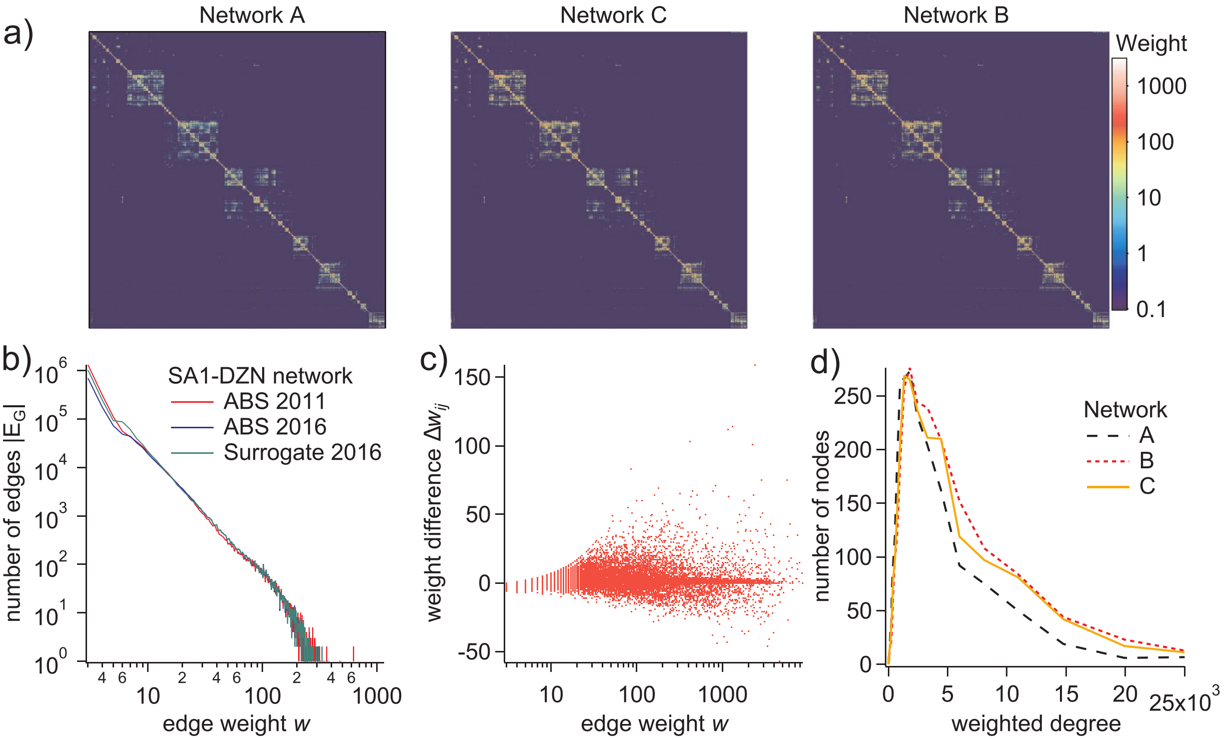

Fig. 6 Validation of the surrogate network. (**a**) Color plots of the SA2 → SA2 adjacency matrix from the aggregate of the original SA1 → DZN network A, the aggregated surrogate C, and the ABS-provided SA2 → SA2 network B. The SA2 regions are somewhat spatially ordered such that the different states, in particular the larger urban areas, are clustered around the diagonal. (**b**) Weight distributions for the networks R, H and S. (**c**) Weight difference, Δw_{ij}, as a function of $w(E_{AB}, B)$, demonstrate improved quantitative agreement (compare to Fig. 3b). (**d**) Distributions of node degree strength (total incident edge weight) for networks A, B, and C.

weights incident on each node), for networks A, B, and C (Fig. 6d). The distribution corresponding to the aggregated surrogate network more closely matches that of the raw SA2 → SA2 network.

We further quantify the similarity between our amalgamated surrogate (C) and the ground-truth network (the edges in network B that also exist in network A), by calculating the mean-squared error (MSE) in the weights over all UR → POW pairs in E_{AB}. Here, we compute the MSE over the edge weight sets

Network pair	*B, A*	*B, C*	*A, C*
D correlation (*r*)	0.9821	0.9996	0.9828

Table 3. 2D correlation coefficients computed according to Eq. 13, between the aggregated and ABS-provided SA2 → SA2 networks. Network *C* is the aggregated SA1 → DZN surrogate, network *A* is the aggregate of the SA1 → DZN network provided by ABS, and network *B* is the SA2 → SA2 network provided by ABS.

Network pair	*B, A*	*B, C*	*A, C*
MSE	62.51	0.27	60.93

Table 4. MSE between the overlapping portions of the aggregated and ABS-provided SA2 → SA2 networks computed according to Eq. 16.

Network	A^*	C^*	B^*	B
Shortest path	0.157	0.118	0.099	0.095
Clustering coefficient ($\times 10^{-3}$)	1.97	2.95	3.11	1.51

Table 5. Average weighted network statistics. The networks marked with an asterisk (*) contain only edges appearing in E_{AB} that is, they represent the overlapping portions of the networks. [Note: inclusion of the edges unique to network *B* quantified in Fig. 3c, produces a dramatic reduction in the network's clustering coefficient, which is intuitive given the relatively low weights of these edges and our definition of the weighted clustering coefficient (Eq. 17)].

$$\alpha = w(E_{AB}, B), \tag{14}$$

and

$$\beta = w(E_{AB}, C) \text{ or, } \beta = w(E_{AB}, A), \tag{15}$$

as

$$\mathrm{MSE}(\alpha, \beta) = \frac{1}{|E_{AB}|} \sum_{e_{ij} \in E_{AB}} \left[\alpha_{ij} - \beta_{ij}\right]^2, \tag{16}$$

where subscripts *ij* indicate specific UR → POW pairs. This quantity provides an estimate of how much our algorithm rectified the discrepancies between SA2 → SA2 edges, given our conservative choice not to add edges to the overlapping set E_{AB}. The results are shown in Table 4 below, and indicate a significant quantitative improvement, as expected from comparison between Figs 3b and 6c.

To evaluate the improvement in the structural properties of the surrogate relative to the as-provided network we analyzed two key network measures for the common components of the networks *A* and *B*. The first is simply the average shortest path between nodes, as computed by applying Dijkstra's shortest-path algorithm to the weighted networks, interpreting edge weight as inverse distance. The second is a version of the clustering coefficient adapted to weighted networks[28] that defines the weighted clustering coefficient for a node *i* by evaluating the fraction of its neighbors *j* and *k* that share connections, weighted based on the relative weights of the edges connecting the triangle, as

$$C_i = \frac{2}{k_i(k_i - 1)} \sum_{j,k} \left(\hat{w}_{ij} \hat{w}_{jk} \hat{w}_{ki}\right)^{1/3}, \tag{17}$$

and reports the average of this quantity over all nodes in the network. Here, the weights of nodes in a triangular cluster are scaled by the largest weight in the network $\hat{w}_{ij} = w\left(\{e_{ij}\}\right) / \max(w(E))$, and k_v is the degree of node *v*.

These network statistics are shown in Table 5 and indicate an improved correspondence between the network properties of the overlapping sets $w(E_{AB}, C)$ and $w(E_{AB}, B)$, as compared to the aggregate of the original network $w(E_{AB}, A)$. These edge sets correspond to the networks labeled as C^*, B^*, and A^* in Table 5, respectively. The improvement is apparent in that the difference between C^* and B^* is smaller than the difference between A^* and B^*.

The number of commuters in the surrogate network is 9,336,333 constituting a 25% increase in the commuter population as compared to the aggregated ABS-provided SA1 → DZN network. Our procedure added nearly half a million new SA1 to DZN edges. The increase in correlation and closer network statistics at the SA2 scale, as well as the edge-wise decrease in mean-squared error indicates both a quantitative and structural improvement over the original dataset provided by the ABS.

Network	*C*	C1	C2	C3
Shortest path	0.119	0.113	0.116	0.119
Clustering coefficient	2.70	2.70	2.65	2.70

Table 6. The weighted network statistics for additional surrogate data sets.

Network	C1	C2	C3
MSE	0.142	0.153	0.148
D correlation	1.000	1.000	1.000

Table 7. The MSE and 2D correlation between the chosen surrogate, C, and additional instances of surrogate networks aggregated to SA2 → SA2 scale.

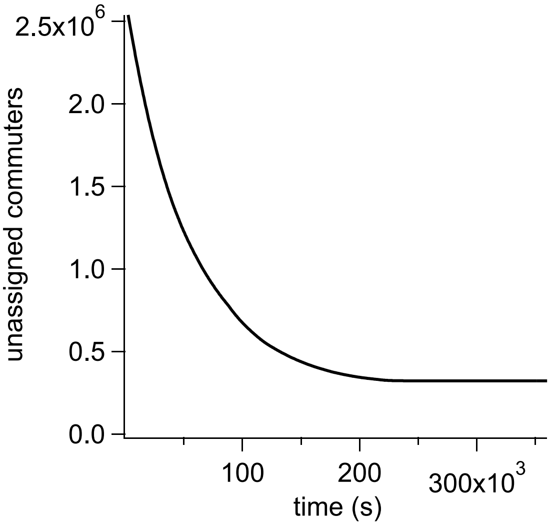

Fig. 7 Algorithm convergence. The number of unassigned commuters as a function of time while assigning commuter weights to the new SA1 → DZN edges. This figure corresponds to running the script 'create_surrogate.m'[27] for 100 hours.

The surrogate network proffered here represents a significant improvement over the original SA1 partitioned commuter mobility network. It reconstructs the population and network statistics of the less perturbed SA2-level network by adding additional SA1 → DZN connections that have been lost to the ABS privacy protocol. Access to the surrogate network, and our example of a method for recovering data on high resolution, anonymized networks is useful for the computational modeling of diffusion and transport phenomena in various disciplines that rely on high-fidelity survey data. The redistribution of ABS data is protected under Creative Commons licensing.

Network statistics for different instantiations. The process of generating the surrogate networks is stochastic. However, the constraints placed on the new edge generation leads to very consistent surrogate network statistics across instantiations. This is evident in comparing the network statistics of the surrogate network analyzed here with several additional instantiations. These are shown in Table 6.

Likewise, the MSE and 2D correlation demonstrate an excellent agreement between the specific surrogate network used for our study, and additional generated surrogates. These are shown in Table 7.

Convergence. The process of building the new edges e^* from the sample edge distributions is the most time consuming part of creating the surrogate networks. Each run generating a surrogate network was given 100 hours to reach the end-point criteria, however a small proportion of commuters remain impossible to assign, as the larger candidate edges become disallowed by the algorithm's constraints. Figure 7 shows the number of unassigned commuters as a function of time when placing the new edges. As edges are added, the constraints of SA1 population, DZN population, and SA2 → SA2 edge weights reduce the likelihood of finding a suitable sample. This results in convergence on a non-zero number of unassigned commuters.

Usage Notes

The MATLAB script 'creating_surrogate.m', available in the online repository[27], implements the method outlined in this paper. The inputs required for this script are located in the repository file 'inputs.mat'. This workspace includes:

- 2016 SA1-DZN commuter network (R),
- 2011 SA1-DZN commuter network (H),
- 2011 SA1 UR populations,
- 2016 SA1 employed residents (N_X),
- 2016 DZN employees (N_Y),
- 2016 SA2-DZN ABS network (Γ),
- SA2-SA2 network accumulated from R (A),
- SA2-SA2 ABS network (B).

Using this script first produces the commuter residential distribution based on the 2011 census data, then a list of possible SA1 edges (M) using the residential distribution, and finally assigns them to DZN partitions, creating e^* samples. These are then combined with the existing edges of network R to create the surrogate network S. A complete description of each network and the file header information is located in the corresponding'README.txt'. The data format is simply a table of edges, the first column corresponding to the SA1 label, the second column corresponding to the DZN label, and the third column giving the number of commuters assigned to the pair.

Code Availability

The custom code used to generate the surrogate network *via* the method outlined in this text was run on MATLAB version R2017b. The script and the required inputs can be accessed on the online repository[27], along with usage notes and descriptions of relevant parameters.

References

1. Yu, F. & James, W. J. High-resolution reconstruction of the United States human population distribution, 1790 to 2010. *Sci. Data* **5**, 180067 (2018).
2. Eubank, S. *et al.* Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184 (2004).
3. Longini, I. M. *et al.* Containing Pandemic Influenza at the Source. *Science* **309**, 1083–1087 (2005).
4. Germann, T. C., Kadau, K., Longini, I. M. & Macken, C. A. Mitigation strategies for pandemic influenza in the United States. *PNAS* **103**, 5935–5940 (2006).
5. Cliff, O. *et al.* Investigating spatiotemporal dynamics and synchrony of influenza epidemics in Australia: an agent-based modelling approach. *Simulat. Model. Pract. Theor* **87**, 412–431 (2018).
6. Wang, Z. *et al.* Statistical physics of vaccination. *Phys. Rep* **664**, 1–113 (2016).
7. Farmer, D. J. & Foley, D. The economy needs agent-based modelling. *Nature* **460**, 685–686 (2009).
8. D'Alelio, D., Libralato, S., Wyatt, T. & d'Alcalà, M. R. Ecological-network models link diversity, structure and function in the plankton food-web. *Sci. Rep* **6**, 21806 (2016).
9. Einav, L. & Levin, J. Economics in the age of big data. *Science* **346**, 1243089 (2014).
10. Lee, J. Y. L., Brown, J. J. & Ryan, L. M. Sufficiency revisited: rethinking statistical algorithms in the big data era. *Am. Stat* **71**, 202–208 (2017).
11. Coull, S. E., Monrose, F., Reiter, M. K. & Bailey, M. The challenges of effectively anonymizing network data. In *2009 Cybersecurity Applications & Technology Conference for Homeland Security* 230–236 (IEEE, 2009).
12. Wooton J. & Fraser B. A review of confidentiality protections for statistical tables, with special reference to the differencing problem. *Australian Bureau of Statistics Methodology Report* ABS Catalogue No. 1352.0.55.072 (2007).
13. Kugler, T. A. & Fitch, C. A. Interoperable and accessible census and survey data from IPUMS. *Sci. Data* **5**, 180007 (2018).
14. Australian Bureau of Statistics *TableBuilder*, http://www.abs.gov.au/websitedbs/D3310114.nsf/Home/2016%20TableBuilder/ (2018)
15. Rogers, D. J. & Cegielski, W. H. Opinion: Building a better past with the help of agent-based modeling. *PNAS* **114**, 12841–12844 (2017).
16. Australian Bureau of Statistics *Australian Statistical Geography Standard (ASGS): Correspondences, July 2011* ABS Catalogue No. 1270.0.55.006 (2013).
17. Coull, S. E., Narayanan, A. & Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE symposium on security and privacy* 111–125 (IEEE, 2008).
18. Sweeney, L. K-anonymity: A model for protecting privacy. *Int. J. Uncaertain. Fuzz* **10**, 557–570 (2002).
19. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **8**, 1000167 (2008).
20. Fraser, B. & Wooten, J. A proposed method for confidentialising tabular output to protect against differencing. *Monographs of Official Statistics: Work Session on Statistical Data Confidentiality* 299–302 (2005).
21. Leaver, V. Implementing a method for automatically protecting user-defined Census tables. *Joint ECE/Eurostat Worksession on Statistical Confidentiality in Bilbao, December 2009* (2009).
22. Wooton, J. Measuring and Correcting for Information Loss in Confidentialised Census Counts. *Australian Bureau of Statistics Research Paper* ABS Catalogue No. 1352.0.55.083 (2007).
23. Zachreson, C. *et al.* Urbanization affects peak timing, prevalence, and bimodality of influenza pandemics in Australia: Results of a census-calibrated model *Science Advances* **4**(12), eaau5294 (2018).
24. Harding, N., Nigmatullin, R. & Prokopenko, M. Thermodynamic efficiency of contagions: a statistical mechanical analysis of the SIS epidemic model. *Interface Focus* **8**, 20180036 (2018).
25. Piraveenan, M., Prokopenko, M. & Zomaya, A. Y. Information-Cloning of Scale-Free Networks. *Advances in Artificial Life* 925–935 (2007).
26. Piraveenan, M., Prokopenko, M. & Zomaya, A. Y. Assortativeness and information in scale-free networks. *The European Physical Journal B* **67**, 291–300 (2009).
27. Fair, K. M., Zachreson, C. & Prokopenko, M. Creating a surrogate commuter network from Australian Bureau of Statistics census data. *Zenodo*. https://doi.org/10.5281/zenodo.2578459 (2018).
28. Onnela, J. P., Saramäki, J., Kertész, J. & Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E* **71**, 065103 (2005).

Acknowledgements

We acknowledge the Australian Bureau of Statistic (ABS) for providing all of the raw data as well as general advice in regards to the nature of their perturbation procedures. The Authors were supported through the Australian Research Council Discovery Project DP160102742.

Author Contributions

K.F., C.Z. and M.P. designed the research; K.F. and C.Z. designed the algorithm; K.F. implemented the algorithm code; C.Z. and K.F. designed the validation strategy; K.F. performed data analysis for validation; C.Z., K.F. and M.P. composed the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.