Journal of Geographical Systems (2021) 23:97–114
https://doi.org/10.1007/s10109-020-00340-4

ORIGINAL ARTICLE

The network boundary specification problem in the global and world city research: investigation of the reliability of empirical results from sampled networks

Vladimír Pažitka[1] · **Dariusz Wójcik**[2]

Received: 18 June 2020 / Accepted: 29 October 2020 / Published online: 22 November 2020

Abstract

Despite the well-known dependence of vertex and network structural parameters on network boundary specification employed by researchers, there has so far been effectively no discussion of this methodological caveat in the global and world city literature. Given the reliance of empirical studies of urban networks on the sampling of underlying actors that form these networks by their interactions, we consider it of key importance to examine the dependence of network centralities of cities on network boundary specification. We consider three distinctive modelling approaches based on: (a) office networks, (b) ownership ties and (c) inter-organisational projects. Our results indicate that city network centralities obtained from sampled networks are highly consistent with those obtained from whole network analysis for samples featuring as little as 4% (office networks), 10% (ownership ties) and 25% (inter-organisational projects) of the underlying actors.

Keywords Network boundary specification problem · Interlocking world city network model (IWCNM) · Inter-organisational project approach (IOPA) · Network sampling · World city network

JEL Classification F30 · F65 · G24 · L10 · R10

✉ Vladimír Pažitka
vladimir.pazitka@ouce.ox.ac.uk

Dariusz Wójcik
dariusz.wojcik@spc.ox.ac.uk

[1] School of Geography and the Environment, University of Oxford, South Parks Road, Oxford OX1 3QY, UK

[2] School of Geography and the Environment, St. Peter's College, University of Oxford, South Parks Road, Oxford OX1 3QY, UK

1 Introduction

The scholarship concerned with studying urban networks through the lenses of corporate networks grounded conceptually in the seminal works of Jacobs (1969), Friedmann (1986), Sassen (2001), Castells (2010) and Taylor and Derudder (2016) has crystalized in the form of three distinctive empirical approaches. These can be broadly divided into those following (1) Taylor's (2001) interlocking world city network model (IWCNM) based on office location data of advanced producer services (APS) firms, (2) ownership linkages among parent companies and subsidiaries of multinational enterprises (Alderson and Beckfield 2004) and (3) inter-organisational projects (Pažitka et al. 2019). We refer to them shortly as the (1) office networks, (2) ownership ties and (3) inter-organisational project approach (IOPA).

The broad division of the existing approaches described here stems primarily from either explicit or implicit choices of the respective researchers regarding the network boundary specification underlying their research design. Network boundary specification in this context refers to the choices researchers make regarding what is and what is not part of the network under study. This problem can be divided into two parts—(1) network ties (edges) and (2) actors (vertices). As a matter of empirical feasibility, researchers generally impose restrictions on the inclusion of both the network ties and actors. Unlike in studies of statistically independent observations, network data by definition includes dependencies among observations and for this reason the network boundary specification problem is of fundamental importance to the validity, robustness and scientific value of studies that focus on network structural properties (Laumann et al. 1989). The established practice in studies of urban networks is to derive the centrality of cities in urban networks by aggregating the network connections of individual organisations located in these cities. Consequently, the implications of the network boundary specification problem are just as relevant to this literature and its conclusions.

In all the empirical studies of urban networks that we consulted, the network boundaries are clearly defined both in terms of the specification of network ties considered as well as actors included (Alderson and Beckfield 2004; Pažitka et al. 2019; Taylor and Derudder 2016). Specification of network ties is generally dictated by the nature of the phenomenon under consideration and availability of data for its empirical observation. Actor inclusion is typically restricted by actors' importance or geographical area of interest. Interestingly, none of the studies of urban networks that we surveyed explicitly discuss the network boundary specification problem and its consequences or reference methodological studies from social network analysis that consider this problem. The primary research objective of this paper therefore is to investigate the reliability of city network centralities obtained from sampled networks. We are specifically interested in the following research question—What sampling percentages are required to obtain city network centralities consistent with those from a whole network analysis?

To allow for a comparable analysis across the three approaches for modelling urban networks outlined earlier, we identify all the securities firms involved in

any of the syndicated deals available in Dealogic Equity Capital Market (ECM) and Debt Capital Market (DCM) databases for 2015, representing collectively 2192 legal entities. We then collect data on their involvement in syndicated deals (Pažitka et al. 2019), office locations (Taylor 2001) and parent-subsidiary ownership ties (Alderson and Beckfield 2004), allowing us to construct three different types of networks for the same set of underlying actors. We then sample from these three distinctive whole networks and compare results across sampled and whole networks, separately for each modelling approach, at sampling percentages ranging from 1 to 95% of the underlying actors. Given the comprehensive nature of Dealogic databases, we treat these urban networks formed by the activities of securities firms as being complete (whole networks), when all the underlying securities firms identified in Dealogic ECM and DCM databases have been included.

Our results indicate that city network centralities obtained from sampled networks converge on those obtained from whole networks as sampling percentage increases. We achieve reliable results with samples ranging from 4% of the underlying actors for Taylor's (2001) IWCNM, 10% for Alderson and Beckfield's (2004) model, to 25% for the IOPA (Pažitka et al. 2019), when we sample the biggest companies first, rather than drawing a random sample. These results contrast with studies of network boundary specification and network sampling in social networks, which suggest that commonly used measures of network centrality become inconsistent with their true values, unless the majority of the underlying actors is sampled (Costenbader and Valente, 2003). We attribute this difference to the specific nature of urban networks formed by APS and particularly the fact that the majority of ties in these networks are formed by the largest and most connected APS firms, while a large number of small firms contribute relatively little to the network centrality of cities, as they often operate from a single location, have no subsidiaries in other cities and seldom engage in inter-organisational projects. This is confirmed by our results based on random sampling, which indicate that much higher sampling percentages—55% (office locations), 100% (ownership ties) and 90% (IOPA)—are needed in order to obtain reliable results. To summarize, researchers can obtain reliable city network centralities in studies of urban networks, provided that sufficient proportion of the most connected firms are included.

The rest of this paper is organized as follows. We begin with reviewing relevant literature on urban networks and contributions from social network analysis related to the network boundary specification problem. The research design section details the methodological basis of the three modelling approaches considered, outlines the methods used for comparing results across sampled and whole networks and presents our dataset. In the results section, we present the results derived across (1) office networks (Taylor 2001), (2) ownership ties (Alderson and Beckfield 2004) and (3) IOPA (Pažitka et al. 2019) and identify the sampling percentages required to obtain reliable city network centralities from each approach. In the concluding section, we reflect on the implications of the network boundary specification problem for literature on global and world cities and make suggestions for future research.

2 The network boundary specification problem in studies of urban networks

2.1 Global and world city research

The global and world city research as well as the wider literature investigating urban networks through the lens of corporate networks rests on the theoretical foundations laid out in the works of Jacobs (1969), Friedmann (1986), Sassen (2001), Castells (2010) and Taylor and Derudder (2016). Friedmann's (1986) *world cities* are conceptualized as centres of power and their world-cityness is examined empirically in studies of ownership links (Alderson and Beckfield 2004). In contrast, Sassen's (2001) *global cities* are service centres hosting advanced producer services (APS) firms' complexes. The primary function of APS firms is to deliver services, such as underwriting of securities, accounting and auditing, management consultancy, marketing and legal services to other businesses. These types of services are designed to make it feasible and economical for owners of financial capital to exercise their economic power, protect their interests and monitor companies that they have invested in at an unprecedented scale. In addition, APS allow corporate boards to manage highly complex organisations, gain access to necessary knowledge and skills, without having to keep all the necessary expertise inhouse. Taylor's (2001) IWCNM has become the most widely used empirical model for studying network ties of Sassen's (2001) global cities. This literature naturally has its critics, who point out methodological weaknesses in the prevailing IWCNM approach (Nordlund 2004; Pažitka et al. 2019) and scrutinize some of the theoretical underpinnings of this literature (Robinson 2005; Smith and Doel 2011). In turn, van Meeteren et al. (2016) engage with many of these critics and Pažitka et al. (2019) offer an alternative modelling approach that overcomes some of IWCNM's limitations.

In this literature review we take a slightly different approach from the previous reviews of research on urban networks (Derudder 2006; Liu et al. 2013). We restrict our attention to studies of corporate networks, given that telecommunication or physical transportation networks are not conceptually strongly grounded in the associated global and world cities literature. As a way of categorizing empirical studies, we focus on the network boundary specification restrictions imposed by researchers. Generally, the research design of every empirical network analysis needs to address two fundamental restrictions on the boundaries of the studied network—the types of network ties (edges) considered and actors (vertices) included.

Taylor (2001) specifies the world city network (WCN) formed by interlocking offices of transnational APS firms. This means that in terms of the network boundary specification, the empirical studies that follow this research design restrict themselves to studying network ties formed by companies with offices

located across city-dyads and impose restrictions on the inclusion of actors based on their industry, importance or geographical location (Derudder et al. 2010; Neal 2008; Sigler and Martinus 2017). Taylor et al. (2002) study a sample of leading APS firms,[1] by selecting those with at least 15 offices, including at least one in Europe, Northern America and Asia–Pacific. They sample those firms with publicly available data on their office locations and restrict themselves to 100 firms to keep the data collection process manageable. Related studies typically use similar sample sizes. For example, Beaverstock et al. (2000) use a sample of leading 74 APS firms, Neal (2008) uses a sample of 100 APS firms, Taylor and Aranya (2008) use samples of 100 (2000) and 80 (2004) APS firms. Derudder et al. (2010), Derudder and Taylor (2016) use sample of 175 leading APS firms made publicly available by the GaWC research group. Due to specifications of their research design, Liu et al. (2013) use a smaller sample of 53 APS firms, a subset of the bigger samples used in other GaWC studies. In all but one of the cases mentioned above APS firms from the following five broad industrial categories are sampled—financial services, accountancy, advertising, law, and management consultancy. In contrast to these studies with global scope, others also impose explicit geographical restrictions on their samples. Sigler and Martinus (2017) draw a sample of 1840 companies listed on the Australian Securities Exchange.

While the focus of the above studies is on studying urban networks formed by offices of organisations, a related stream of literature focuses on power relations by observing ownership linkages among parent companies and subsidiaries of multinational enterprises (MNEs). The studies of Alderson and Beckfield (2004) and Alderson et al. (2010) use global samples of 500 leading MNEs. Both studies consider network connections to be directed from the parent to the subsidiary and only count edges in this direction. In a follow up study, Wall and van der Knaap (2011) analyse a network of 100 leading MNEs to examine the power structure of the WCN. More recently, Rozenblat et al. (2017) apply this modelling approach to a sample of the top 3000 MNEs, controlling a portfolio of over 800,000 subsidiaries.

Finally, we also consider approaches based on inter-organisational projects as the elementary building blocks of the WCN. Pan et al. (2017) have pioneered the use of inter-firm service provision relationships as an alternative to intra-firm ties, which formed the basis of earlier world city network studies (Taylor and Derudder, 2016). In contrast to studies of office networks, this stream of research builds on the premise that urban networks can be mapped out by tracing ties formed by inter-firm collaborations on inter-organisational projects, which relate to the provisions of APS. Pan et al. (2017) derive such inter-firm ties from the joint provision of financial, accounting and legal services in initial public offerings (IPOs) by securities underwriters, accounting and law firms. Their study utilizes a sample of IPOs listed on the Shanghai Stock Exchange and Shenzhen Stock Exchange from 2004 to 2014 covering 1318 issuers, 113 securities firms, 99 accounting firms and 165 law firms. In a subsequent study Pan et al. (2018) extend this analysis to 2296 IPOs of Chinese domestic firms on Shanghai and Shenzhen stock exchanges for the period

[1] Financial services, accountancy, law, advertising, and management consultancy.

1993–2014. Pažitka et al. (2019) develop an alternative model of urban networks based on the ties formed by the membership of investment banks in underwriting syndicates and advance the argument on the appropriateness of the use of inter-firm service relationships for modelling urban networks by utilizing the construct validity framework (Cronbach and Meehl 1955; Messick 1995). The study of Pažitka et al. (2019) draws a global sample of 161,114 underwriting syndicates to analyse WCN and its evolution for the 2000–2015 period.

2.2 The network boundary specification problem

The network boundary specification is recognized by researchers in social network analysis as one of the most important decisions in studies using network data. It entails decisions made by researchers regarding the specification of network ties considered and inclusion rules for actors in the studied network. While the specification of ties depends primarily on the phenomenon under study and availability of measurable interactions, the choice of actors to be included should reflect the true boundaries of the network. Researchers have adopted two distinct approaches to specifying boundaries of networks—(1) realist and (2) nominalist approach. Realist approach requires studied subjects to self-identify the boundaries of the network by reporting their network connections to other actors. In contrast, nominalist approach relies on a researcher to draw the network boundaries. The realist approach is generally better suited to studies involving fieldwork and primary data collection, while studies relying on secondary data generally adopt the nominalist approach (Laumann et al. 1989).

Any restrictions on the inclusion of actors or ties imposed by researchers that result in an incomplete sociocentric[2] image of the network lead to a missing data problem. This is a much more serious concern in network analysis than in analysis of statistically independent observations, because observed vertex and network structural parameters rely potentially on all observed vertices and edges. Consequently, even if a random sample of vertices is drawn, it cannot be guaranteed that this will lead to a sampled network with structural properties that are representative of the whole network. For a researcher following a nominalist approach, opting to identify the network boundary herself rather than relying on studied subjects to identify boundaries of their network, there are two principal approaches available—(1) sociocentric and (2) egocentric approach (Doreian and Woodard 1994). Following a sociocentric approach is feasible, if there is a prior knowledge on the full set of actors belonging to a given network. In situations when this is not the case, it is possible to pursue the egocentric approach and identify the approximate boundaries of a network by repeated rounds of snowball sampling (Stumpf et al. 2005).

The methodological research on network boundary specification, network sampling and missing data in networks, which represent three facets of the same problem, has focused primarily on the measurement error, bias and reliability of vertex

[2] Sample that includes all vertices and edges of a network.

and network structural attributes in one-mode networks resulting from randomly missing data. In a pioneering study, Galaskiewicz (1991) considered the effect of network sampling on the vertex level in-degree centrality measures in two empirical networks constituting a broad range of private, public and third sector organisations. The results of these simulations indicate that the variance of measurement errors decreases with sampling percentage. Bolland (1988) found that the Pearson correlation between vertex centralities in unperturbed networks and those with simulated errors in the data decreases approximately linearly with the number of simulated errors. In a related study, Johnson et al. (1989) consider the effect of initial sample size, number of choices and number of sampling waves on the precision of vertex in-degree centralities. Their results indicate there is a trade-off between these parameters, when attempting to decrease measurement error. These findings are corroborated by a broader simulation study of Borgatti et al. (2006), which shows that the precision of vertex centralities declines smoothly and linearly with the fraction of missing data.

In a large-scale study based on 59 empirical networks Costenbader and Valente (2003) show that different centrality measures vary greatly in their correlation with their true values, when only a fraction of a network is sampled. Their results indicate that the centrality measures commonly used by social scientists are generally reliable for sample sizes of at least 80% of vertices and their respective edges, while effectively none are reliable for sample sizes of less than 20%. Zemljič and Hlebec (2005) consider the effect of measurement errors associated with network surveys and their impact on the reliability of vertex centrality measures. Local measures are shown to be more robust than global measures and network density mitigates the impact of measurement error on reliability of centrality measures. In contrast to much of the related literature, Kossinets (2006) considers the effect of three different sources of missing data—network boundary specification, survey non-response and censoring by vertex degree on network structural measures of bipartite graphs. Network boundary specification and censoring by vertex degree are shown to have the most severe impact on estimates of network structural measures, while survey non-response is of much smaller relevance, especially if bipartite graphs feature high degree of redundancy.

It has been shown across the wider literature on network sampling, missing data and network boundary specification that these matters should be given full consideration by researchers investigating vertex and network structural parameters. Network boundary specification is also relevant to studies using blockmodels (Žnidaršič et al. 2012), Exponential Random Graph Models (Valente et al. 2013) or Stochastic Actor-Oriented Models (Leszczensky and Pink 2015). Apart from the most simplistic of scenarios[3] it seems to be difficult to predict the consequences of network boundary restrictions that lead to important omissions in the network data and even more so to correct for it. For this reason, it seems that especially in the investigation of structural properties of networks it is important to obtain a reasonably complete dataset capturing both vertices and edges of the network accurately.

[3] Random networks and random patterns of missing data.

3 Research design

3.1 Modelling approaches in urban network analysis

Taylor's (2001) interlocking world city network model (IWCNM) derives estimates of urban network connectivity from the spatial distribution of offices of advanced producer services (APS) firms. This is operationalized by using a matrix of 'service values' V, with rows representing cities and columns representing firms. Elements of $V - v_{i,j}$ are customarily coded on a 0 to 5 scale,[4] with 0 for non-presence in a city and 5 for a global head office. The individual network connectivity scores $r_{i,j}$ representing the strength of connection between two offices of a particular firm are given by the product of service values for the given office-dyad (Eq. 1). Estimates of flows for city-dyads are then given by the sums of connectivity scores for office-dyads that connect a given city-dyad (Eq. 2). Finally, to calculate a measure of status of city a within the WCN (N_a), Taylor (2001) proposed to sum the relational elements $r_{i,j}$ for each city a, thus giving us an aggregate measure of connectivity between city a and all other cities (Eq. 3). N_a can be therefore interpreted as a weighted degree centrality of cities, produced by summing network connectivity across all firms located in that city and counting only those connections that connect offices of firms in city a with those outside of it.

$$r_{ab,j} = v_{aj} \cdot v_{bj} \tag{1}$$

$$r_{ab} = \sum_{j} r_{ab,j} \tag{2}$$

$$N_a = \sum_{i} r_{ai} \quad a \neq i \tag{3}$$

In contrast, Alderson and Beckfield (2004) base their approach on ownership ties among parent companies and subsidiaries. They count only directed network connections from the city of the parent to the city of the subsidiary. City network centralities in Alderson and Beckfield (2004) approach are given by the number of subsidiaries outside of city a owned and controlled by parent companies headquartered in city a.

Finally, the inter-organisational project approach (Pažitka et al. 2019) relies on data on syndicated deals, in the form of a firm-deal affiliation matrix. Firms are recorded at the lowest available subsidiary level reported in Dealogic ECM and DCM databases in relation to each deal in order to maximize the geographical accuracy of allocating deals to cities. This is then converted to a firm adjacency matrix by the projection function below:

[4] Due to the limitations of our data, we only code offices on a three-point scale: 5 = global HQ, 4 = national HQ, 3 = regular office.

$$B = D\left(D^T\right); b_{i,j} = 0 \; if \; i = j \tag{4}$$

where D is a bank—deal affiliation matrix, D^T is a transpose of D, B is a weighted bank adjacency matrix and $b_{i,j}$ are frequencies of co-membership of firm-dyads in syndicated deals.

To obtain intercity network ties, we convert the bank level adjacency matrix B into an edge-list of bank-dyads and include frequencies of co-syndication as weights for network ties b_{ij}. We then add data on the location of operational headquarters of bank subsidiaries to allow us to sum the bank– bank edge weights $b_{ij,vw}$ by city-dyads vw (Eq. 5). We define operational headquarters as the publicly known head office of the company, where most of its top-level management team and key personnel is normally based. This is in contrast with financial headquarters, which is often located separately for tax or other purposes. Finally, we convert it to a weighted city adjacency matrix C with elements c_{vw}, which represent intercity network ties formed by co-syndication of banks' subsidiaries located in cities v and w.

$$c_{vw} = \sum_{vw} b_{ij,vw} \tag{5}$$

We then use the methodology for calculating group degree centrality developed by Everett and Borgatti (1999) to derive the network centrality of cities, which is defined as the sum of non-redundant network connections between firms in city a and those outside of city a. The non-redundancy condition ensures that network connections are not double counted and at most one connection to every firm outside of city a is counted towards its group degree centrality score. We standardize all city network centralities by the maximum available value, separately for each approach.

3.2 Sampled vs. whole networks

We first obtain city network centralities derived from three approaches for modelling urban network connectivity—office network (Taylor 2001), ownership ties (Alderson and Beckfield 2004) and inter-organisational project approach (IOPA) (Pažitka et al. 2019). We then compare the results derived from networks with sampling percentages from 1 per cent to 95 per cent of the underlying actors to those obtained from a whole network analysis[5] for each approach separately. The underlying actors in this instance are underwriters of equity and debt securities. We sample at the level of parent companies, because this sampling method can be meaningfully applied to all three modelling approaches and allows for like with like comparison of the results. Finally, we identify minimum required sampling percentages, that yield network centralities, which are highly consistent with those from a whole network.

We consider both random sampling, which has been commonly used in studies of network sampling and network boundary specification (Borgatti et al. 2006; Costenbader and Valente 2003; Galaskiewicz 1991; Stumpf et al. 2005) as well as sampling

[5] We use the term 'whole network' to say that all the underlying actors were included.

from the top of a list of firms ordered by size (top-down sampling), which has been widely employed in the global and world city literature (Alderson et al. 2010; Taylor and Derudder 2016). Given the abundance of studies of urban networks in the global and world city literature that consider city network centralities and produce ranking tables, we consider both the raw network centralities and their respective ranking. Finally, to compare city network centralities for sampled and whole networks, we employ the Pearson correlation coefficient and Kullback–Leibler divergence, both of which are widely used in the research of network sampling and network boundary specification (Vedral 2002; Zemljič and Hlebec 2005; Villas Boas et al. 2010).

Pearson correlation coefficient can be used as a measure of linear correlation between two variables. In this instance, we use it to compare city network centralities for the same set of cities obtained from sampled and whole networks. Carmines and Zeller (1979) show that Pearson correlation coefficient can be validly interpreted as a measure of reliability, when applied as a means of comparison of perfect and imperfect measures of the same phenomenon. We therefore use Pearson correlation coefficient as a measure of reliability of city network centralities obtained from sampled networks, by comparing them to their values obtained from whole networks.

As a robustness check, we employ Kullback–Leibler divergence (KLD), an entropy-based measure, which has been widely used to quantify the distance between two probability distributions (Vedral 2002). Villas Boas et al. (2010) show that KLD is well suited for quantifying the perturbation of the network structural parameters, by comparing their probability distributions across unperturbed (whole network) and perturbed (some missing data) states. Sampled networks investigated by us are analogous to networks with missing vertices and therefore KLD is appropriate to test the extent of perturbation caused by various degrees of missing data.[6]

We use both Pearson correlation coefficient and KLD to compare city network centralities obtained from sampled networks[7] with those obtained from the whole network. Given the comprehensive coverage of Dealogic ECM and DCM databases we treat all of the securities companies involved in all of the syndicated deals combined as the full set of actors. To compare city network centralities among sampled and whole networks, we calculate Pearson correlation coefficient and KLD for pairwise combinations of city network centralities obtained at sampling percentages ranging from 1 to 95% and those from the whole network. For random samples of the underlying actors we use a bootstrapping procedure similar to that of Costenbader and Valente (2003) based on 50 random samples at each sampling interval. We then average Pearson correlations and KLD across these samples and report their averages at each sampling interval. We also consider networks obtained by top-down sampling and report Pearson correlations and KLD for each sampling percentage. top-down samples are obtained by sampling parent companies from the top of an ordered list, using number of syndicated deals they were involved in 2015 as the ordering variable.

[6] Sampling fraction of x % in turn means that $100\% - x$ % of parent companies from the whole network are missing.

[7] We use sampling percentages ranging from 1 to 95% of parent companies available in our dataset.

To allow for a simple interpretation of our results and to draw readily implementable recommendations for applied researchers, we set a threshold of $r=0.99$ for Pearson correlation coefficient to assess the reliability of city network centralities from sampled networks. We then identify a minimum sampling percentage at which city network centralities from sampled networks yield $r=0.99$. We interpret this as the minimum sampling percentage required to obtain city network centralities, which are highly consistent with those that would have been obtained from a whole network analysis. We understand that this threshold is in part arbitrary, although by setting it at such a high level, one can be confident that using the minimum sampling percentages identified here for different modelling approaches will yield results, which are only minimally affected by the network boundary specification problem.

3.3 Data

To allow for a whole network analysis, we identify all the syndicated capital market deals in the Dealogic Equity Capital Market (ECM) and Debt Capital Market (DCM) databases for the year 2015. This search yielded 13,666 deals in total—2539 and 11,127 from ECM and DCM databases, respectively. The number of deals varies widely across securities firms and the bulk of the deals feature the largest investment banks. As an example, Morgan Stanley has been involved in 2472 deals, Bank of America Merrill Lynch in 2143 and JPMorgan in 2070 deals.

We then identify all the securities firms involved in these deals, leading to a sample of 272 parent companies with one or more subsidiaries and 1126 independent companies[8] operating out of a single headquarter location. Collectively, the parent companies in our sample control a portfolio of 1066 subsidiaries involved in ECM and DCM deals. Similarly, as in the case of the distribution of deals per firm, the distribution of subsidiaries per parent is also heavily skewed with the largest banks controlling a disproportionately large number of subsidiaries. HSBC controls 29 subsidiaries involved in ECM and DCM deals, Citi and JPMorgan both control 20 subsidiaries and ING and Deutsche Bank control 18 and 17 subsidiaries, respectively. As the next step, we identify the metropolitan areas of headquarters of operations at both subsidiary and parent level. Parent companies in our sample are located across 279 cities globally and the subsidiaries cover an archipelago of 306 cities.

To identify the cities in which securities firms have their offices, we relied on the Dealogic ECM and DCM data, which allowed us to identify the lowest level subsidiaries for each company involved in each capital market deal. Given the organisational structure of the securities industry, securities firms operate legally separate subsidiaries across different countries and occasionally own a portfolio of subsidiaries located in a single country. The location of operational headquarters of these subsidiaries is consequently informative regarding the geographical distribution of the key offices of each parent company. Given the nature of capital market deals and the expertise required for their underwriting, we assume that they are underwritten

[8] We define independent companies as those that do not have a parent company or subsidiaries.

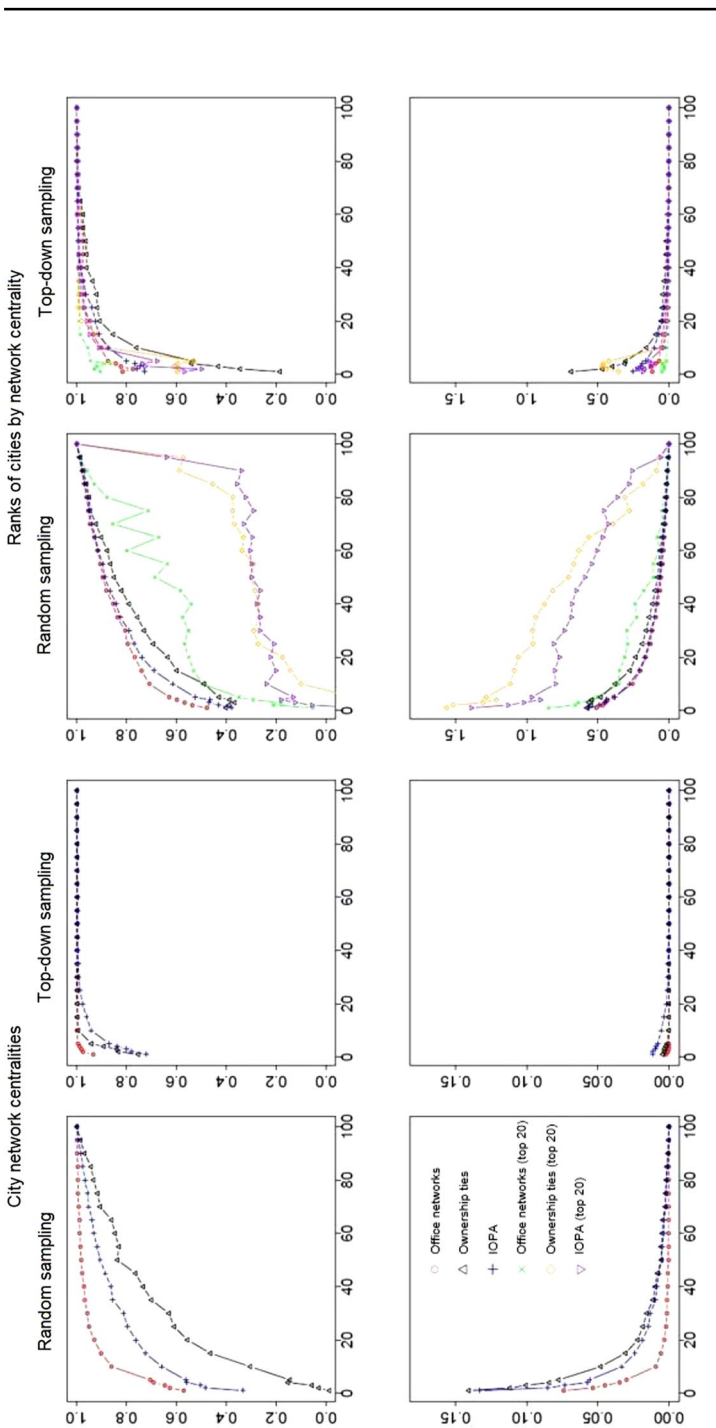

Fig. 1 Comparison of city network centralities across sampled and whole networks. Notes: ρ—Pearson correlation coefficient; KLD—Kullback–Leibler divergence (Vedral 2002); Random sampling—networks produced by randomly sampling investment banks; top-down sampling—networks produced by sampling investment banks from the top of an ordered list (ordered by investment banking revenue); top 20—comparison of ranks of the top 20 cities in the urban network across sampled and whole networks. Source: Authors' analysis based on Dealogic data

by the headquarters of the respective subsidiaries of securities firms and allocate them accordingly. HSBC has been involved in ECM and DCM deals through a network of offices located in 19 cities worldwide, office networks of Citi and Deutsche Bank are located across 13 cities and the office network of BNP Paribas covers 12 metropolitan areas. In terms of geographical distribution, our sample covers 573 securities firms in Americas, 992 in EMEA and 627 in Asia–Pacific.

4 Results

We now compare the city network centralities obtained from the whole network analysis to those from sampled networks. We do this separately for office networks, ownership ties and IOPA. Our results presented in Fig. 1 indicate a convergence following a concave trajectory of city network centralities obtained from all three modelling approaches. Randomly sampled networks require a high percentage of securities firms to be sampled for their results to converge on those obtained from the whole network. As an example, we find that 55% (office networks), 100% (ownership ties) and 90% (IOPA) of underlying actors need to be sampled to obtain reliable ($r=0.99$)[9] city network centralities. In this respect city network centralities based on office networks are relatively more reliable for incomplete random samples of the underlying actors than those obtained using ownership ties or IOPA.

In contrast, samples of as little as 4% (office networks), 10% (ownership ties), 25% (IOPA) of the underlying actors yield results that are just as reliable ($r=0.99$), if we sample from the top of a list of investment banks ordered by size. This notable improvement certainly seems to justify the preference of applied researchers for top-down sampling in studies of urban networks. This is an important finding not only for reducing the cost of data collection, but it also supports the notion that city network centralities can be reliable, even if only a fraction of the underlying actors is sampled.

Word of caution must be however exercised here. This finding is only applicable to situations, where a well-defined ordered list of all relevant actors is available to begin with. If this is not the case and random sampling is applied instead, much higher sampling percentages of the underlying actors are required as illustrated above. In addition, the most important vertices in the network and their respective ranking can be impacted disproportionately, when random sampling is applied. As shown in Fig. 1, if top-down sampling is applied, the results for the top 20 cities are consistent with those obtained from the whole network, even for low sampling percentages of the underlying firms. This is certainly good news for studies of APS firms, MNEs or listed corporations, for which a plethora of ranking tables is available, or they can be constructed from datasets used in empirical research.

In the results reported above we set a threshold of 0.99 for a Pearson correlation coefficient among city network centralities from sampled (incomplete) and whole networks as a cut-off point for assessing reliability. Naturally, this threshold is in part

[9] Pearson correlation coefficient.

Table 1 Robustness tests of the cut-off point of Pearson correlation coefficient for assessing consistency of city network centralities

Pearson correlation coefficient	Office locations		Ownership ties		Inter-organisational projects	
	Random	Top of the list	Random	Top of the list	Random	Top of the list
	Sampling % of the underlying actors					
0.99	55	4	100	10	90	25
0.95	25	2	90	10	70	15
0.9	20	1	70	5	50	10

Pearson correlation coefficient: measures the correlation between city network centralities obtained from sampled networks and those obtained from whole networks, which include the population of the underlying actors. We use it as a measure of consistency between city network centralities from sampled networks and their true values. We use different cut-off points of Pearson correlation coefficient (0.90, 0.95, 0.99) as the cut-off point for consistency to examine, how the minimum sampling percentage for different modelling approaches (office networks, ownership ties, inter-organisational projects) changes in response. Random sampling—networks produced by randomly sampling investment banks; Top-down sampling—networks produced by sampling investment banks from the top of an ordered list (ordered by investment banking revenue)

Source: Authors' analysis of Dealogic data

arbitrary, despite our view that it is prudent to set it at a high level to ensure a high degree of reliability of results in empirical research. In Table 1 below we show, how the required minimum sampling percentages vary, if we lower this threshold to 0.95 and 0.90. Starting with office networks, reducing the threshold for Pearson correlation coefficient from 0.99 to 0.90 reduces the required minimum sampling percentage of the underlying actors from 4 to 1% for top-down sampling and from 55 to 20% for random sampling. We observe quantitatively smaller reductions in the required minimum sampling percentage for ownership ties and IOPA. The required minimum sampling percentage drops from 10 to 5% for top of the top-down sampling and from 100 to 70% for random sampling for the ownership ties approach (Alderson and Beckfield 2004). IOPA requires minimum sampling percentages of 10% for top-down sampling and 50% for randomly sampled networks, when we reduce the required threshold for reliability to $r=0.90$.

5 Conclusions

The purpose of this paper has been to investigate the consistency of city network centralities obtained from sampled networks with those from whole networks. We consider three modelling approaches used in the global and world city research, based on office networks (Taylor 2001), ownership ties (Alderson and Beckfield 2004) and inter-organisational projects (Pažitka et al. 2019). We have

been specifically interested in the following research question—what sampling percentages are required to obtain city network centralities consistent with those from a whole network analysis?

We analyse the reliability of city network centralities by comparing results obtained from networks with sampling percentages ranging from 1 to 95% of the underlying actors with those from a whole network analysis. Despite the argument advanced by Burt (1983), suggesting that sampling anything less than the whole network has a severe impact on network data, our results indicate that it is possible to sample network data and still obtain reliable results on vertex structural attributes. Our results obtained using simulated random samples of actors are broadly consistent with those of Costenbader and Valente (2003), Borgatti et al. (2006) and show that network centralities obtained from sampled networks become more similar to those from the whole network as the sampling percentage increases. However, as we show in our analysis of networks constructed by top-down sampling of the underlying actors, results obtained from sampled networks become highly correlated ($r = 0.99$) with those for the whole network for sampling percentages as low as 4% (office networks), 10% (ownership ties) and 25% (IOPA) of the underlying actors. In case of randomly sampled networks, considerably higher sampling percentages are required to achieve the same level of consistency. Consequently, it seems preferable to use sampling techniques that ensure that the most central actors are not omitted, such as top-down sampling considered here or snowball sampling, which has been shown to have similar properties (Galaskiewicz 1991; Johnson et al. 1989).

In the context of the global and world cities literature, most studies benefited from a prior knowledge regarding the importance or size of their underlying actors and can therefore utilize top-down sampling to their advantage. In order to evaluate the reliability of results of empirical studies, it is crucial to be aware of the sampling percentage of the underlying actors, which has not been routinely reported in existing research (Alderson et al. 2010; Liu et al. 2013; Taylor and Derudder 2016). We therefore suggest that researchers first identify the size of the population of their underlying actors, for example by consulting a comprehensive database such as Bureau van Dijk Orbis. One can then proceed to either build whole networks by including all relevant actors, if feasible, such as in the case of data on syndicated deals available from Dealogic or Thomson Reuters, or data on parent-subsidiary ties available from Bureau van Dijk Orbis. Alternatively, in cases when researchers hand-collect network data, they may wish to instead aim for the minimum sampling percentages identified here and apply top-down sampling, to minimize data collection costs.

The analysis presented here is subject to several limitations. First, we only consider three specific network centrality measures that have been applied in the global and world city literature. There is a plethora of other network structural parameters that have been considered in network studies. The reason we focus on these city network centralities is their relevance to the global and world city literature and their distinctive methodological underpinnings. Second, we consider random sampling and top-down sampling. There are other widely used sampling procedures, such as snowball sampling, which has been widely used in the social network analysis.

We omitted snowball sampling here, given that to the best of our knowledge, it has not been applied in studies of urban networks. Instead, studies of urban networks formed by APS and MNEs almost exclusively rely on sampling from the top of an ordered list, either available in a form of an externally made ranking table, such as Fortune Global 500 used by Alderson et al. (2010) or derived from the data itself, such as the top 500 investment banks by deal value used by (Pažitka et al. 2019). Finally, unlike many empirical studies of urban networks (Sigler and Martinus 2017; Pan et al. 2018), we do not consider multiple industries that we could compare urban networks for. This would certainly be an interesting extension to our analysis.

We hope that our findings and recommendations will be considered by applied researchers, when designing and interpreting studies of urban networks and that others will build on this contribution.

Acknowledgements This research was supported under Australian Research Council's Discovery Projects funding scheme (Project DP160103855), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 681337), and the Hong Kong Research Grants Council (T31-717/12-R). All errors and omissions are the sole responsibility of the authors.

References

Alderson AS, Beckfield J (2004) Power and position in the World City system. Am J Sociol 109(4):811–851

Alderson AS, Beckfield J, Sprague-Jones J (2010) Intercity relations and globalisation: the evolution of the global urban hierarchy, 1981–2007. Urban Stud 47(9):1899–1923

Beaverstock JV, Smith RG, Taylor PJ (2000) World-city network: a new metageography? Ann Assoc Am Geogr 90(1):123–134

Bolland JM (1988) Sorting out centrality: an analysis of the performance of four centrality models in real and simulated networks. Soc Netw 10(3):233–253

Borgatti SP, Carley KM, Krackhardt D (2006) On the robustness of centrality measures under conditions of imperfect data. Soc Netw 28(2):124–136

Burt RS (1983) Applied network analysis: a methodological introduction. Sage, Beverley Hills

Carmines EG, Zeller RA (1979) Reliability and validity assessment, vol 17. Sage, Beverley Hills

Castells M (2010) The rise of the network society. Blackwell Publishing, Malden

Costenbader E, Valente TW (2003) The stability of centrality measures when networks are sampled. Soc Netw 25(4):283–307

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52(4):281–302

Derudder B (2006) On conceptual confusion in empirical analyses of a transnational urban network. Urban Stud 43(11):2027–2046

Derudder B, Taylor PJ (2016) Change in the World City Network, 2000–2012. Prof Geogr 68(4):624–637

Derudder B, Taylor PJ, Ni P, De Vos A, Hoyler M, Hanssens H et al (2010) Pathways of change: shifting connectivities in the world city network, 2000–2008. Urban Stud 47(9):1861–1877

Doreian P, Woodard KL (1994) Defining and locating cores and boundaries of social networks. Soc Netw 16(4):267–293
Everett MG, Borgatti SP (1999) The centrality of groups and classes. J Math Sociol 23(3):181–201
Friedmann J (1986) The World City hypothesis. Dev Change 17(1):69–83
Galaskiewicz J (1991) Estimating point centrality using different network sampling techniques. Soc Netw 13(4):347–386
Jacobs J (1969) The economy of cities. Random House, New York
Johnson JC, Boster JS, Holbert D (1989) Estimating relational attributes from snowball samples through simulation. Soc Netw 11(2):135–158
Kossinets G (2006) Effects of missing data in social networks. Soc Netw 28(3):247–268
Laumann EO, Marsden PV, Prensky D (1989) The boundary specification program in network analysis. In: Freeman LC, Romney AK, White DR (eds) Research methods in social network analysis. Transaction Publishers, Piscataway, pp 61–88
Leszczensky L, Pink S (2015) Ethnic segregation of friendship networks in school: testing a rational-choice argument of differences in ethnic homophily between classroom- and grade-level networks. Soc Netw 42:18–26
Liu X, Derudder B, Liu Y, Witlox F, Shen W (2013) A stochastic actor-based modelling of the evolution of an intercity corporate network. Environ Plan A 45(4):947–966
Messick S (1995) Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol 50(9):741–749
Neal ZP (2008) The duality of world cities and firms: comparing networks, hierarchies, and inequalities in the global economy. Glob Netw 8(1):94–115
Nordlund C (2004) A critical comment on the taylor approach for measuring World City interlock linkages. Geogr Anal 36(3):290–296
Pan F, Bi W, Lenzer J, Zhao S (2017) Mapping urban networks through inter-firm service relationships: the case of China. Urban Stud 54(16):3639–3654
Pan F, Bi W, Liu X, Sigler TJ (2018) Exploring financial centre networks through inter-urban collaboration in high-end financial transactions in China. Reg Stud 54(2):162–172
Pažitka V, Wójcik D, Knight E (2019) Critiquing construct validity in World City network research: moving from office location networks to inter-organizational projects in the modeling of intercity business flows. Geogr Anal. https://doi.org/10.1111/gean.12226
Robinson J (2005) Urban geography: world cities, or a world of cities. Prog Hum Geogr 29(6):757–765
Rozenblat C, Zaidi F, Bellwald A (2017) The multipolar regionalization of cities in multinational firms' networks. Glob Netw 17(2):171–194
Sassen S (2001) The Global City: New York, London, Tokyo, 2nd edn. Princeton University Press, Princeton
Sigler TJ, Martinus K (2017) Extending beyond "world cities" in World City Network (WCN) research: urban positionality and economic linkages through the Australia-based corporate network. Environ Plan A 49(12):2916–2937
Smith RG, Doel MA (2011) Questioning the theoretical basis of current global-city research: structures, networks and actor-networks. Int J Urban Reg Res 35(1):24–39
Stumpf MPH, Wiuf C, May RM (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. Proc Natl Acad Sci 102(12):4221–4224
Taylor PJ (2001) Specification of the World City network. Geogr Anal 33(2):181–194
Taylor PJ, Aranya R (2008) A global "urban roller coaster"? Connectivity changes in the world city network, 2000–2004. Reg Stud 42(1):1–16
Taylor PJ, Derudder B (2016) World City network: a global urban analysis, 2nd edn. Routledge, London
Taylor PJ, Catalano G, Walker DR (2002) Exploratory analysis of the world city network. Urban Stud 39(13):2377–2394
Valente TW, Fujimoto K, Unger JB, Soto DW, Meeker D (2013) Variations in network boundary and type: a study of adolescent peer influences. Soc Netw 35(3):309–316
Van Meeteren M, Derudder B, Bassens D (2016) Can the straw man speak? An engagement with postcolonial critiques of "global cities research". Dialog Hum Geogr 6(3):247–267
Vedral V (2002) The role of relative entropy in quantum information theory. Rev Mod Phys 74(1):197–234
Villas Boas PR, Rodrigues FA, Travieso G, Costa LDF (2010) Sensitivity of complex networks measurements. J Stat Mech: Theory Exp 3:1–19

Wall RS, van der Knaap GA (2011) Sectoral differentiation and network structure within contemporary worldwide corporate networks. Econ Geogr 87(3):267–308
Zemljič B, Hlebec V (2005) Reliability of measures of centrality and prominence. Soc Netw 27(1):73–88
Žnidaršič A, Ferligoj A, Doreian P (2012) Non-response in social networks: the impact of different non-response treatments on the stability of blockmodels. Soc Netw 34(4):438–450