Applied Network Science

RESEARCH Open Access

# Structural studies of the global networks exposed in the Panama papers

Mayank Kejriwal* and Akarsh Dang

*Correspondence: kejriwal@isi.edu
Information Sciences Institute,
University of Southern California,
4676 Admiralty Way, Ste. 1001,
Marina del Rey, California, United
States of America

**Abstract**

In recent history, the Panama Papers have comprised one of the largest and most influential leaks detailing information on offshore entities, company officers and financial (and legal) intermediaries, and has led to a global exposé of corruption and tax evasion. A systematic analysis of this information can provide valuable insights into the structure and properties of these entities and the relations between them. Network science can be applied as a scientific framework for understanding the structure of such relational, heterogeneous datasets at scale. In this article, we use an existing, relational version of the Panama Papers to selectively construct various networks, and then study the properties of the underlying system using well-defined analytical methods from network science, including degree properties, country assortativity analyses, connectivity and single-point network metrics like transitivity and density. We also illustrate significant structural features in these networks by conducting a triad census and exploring the networks' core-periphery structure. Together, these results are used to show that the Panama Papers constitute a distinct class of networks that differ significantly from ordinary social and information networks. We also propose, construct and analyze 'higher-order' networks from the raw data, such as a 'social' network of officers. We confirm that some of these higher-order networks also show significant non-random deviations from expected or typical behavior, including in their degree distributions.

**Keywords:** Panama papers, Offshore finance, Network science, Structural network properties, Motif analysis, Visualization

## Introduction

In 2015, an anonymous source leaked more than 11.5 million documents that detail financial and attorney-client information for more than 214,488 offshore entities. Although the 'Panama Papers', as they are now termed, are now widely associated with investigating corruption, money laundering and tax evasion, the original goal of the (still anonymous) whistleblower, according to a released statement, was to expose income inequality and injustice. Because of the scope of the Panama Papers, the findings capture the impact of globalization as a complex network within a relatively constrained setting. These papers represent an important milestone in the use of data journalism software tools and mobile collaboration. Their impact[1] is already being felt in many governments; according to a recent report, more than 1.2 billion USD in back taxes and penalties have already been collected by governments around the world (Obermayer and Obermaier 2016).

However, precisely because the collection maps out a global system, the Panama Papers also present us with a golden opportunity to study the flow of information[2] between firms, individuals and intermediaries. From a scientific perspective, the Panama Papers represent a complex system, with entities that range from individuals to companies, many of which serve a specific purpose based on where in the world they are based, to a variety of relationships. Studying the *structural properties* of this complex system using applied networks science has the potential to reveal interesting trends about how such systems operate across geographies and economies.

In this paper, we use the publicly available data to selectively construct networks to study the structural properties of such a global, inter-connected system of offshore entities, officers and intermediaries. Specifically, we construct and study eight networks, each of which provides a different lens on the global system represented by the Panama Papers. We describe our methodology in detail in the next few sections, followed by a network-theoretic presentation and analysis of the results that seek to answer some of the questions posed above. Specific contributions are described below.

### Key contributions and findings

Our primary contribution in this article is to formulate a research agenda around deriving insights into the global system represented in the Panama Papers, which contain entities and officers from more than 100+ countries. We posit that, rather than small-scale or qualitative analysis, network science is an appropriate methodology to study such a complex system and to derive structural insights. To our knowledge, a network-theoretic analysis on the currently available Panama Papers has not been done prior to this work. With this core high-level contribution in mind, we list some specific findings and contributions below:

1 We do a detailed structural investigation of entities (including offshore organizations, intermediaries and company officers) in the Panama Papers by constructing a selective set of networks from the raw data and studying structure by computing network-theoretic metrics, including distributional metrics like the degree distributions of the networks, single-point metrics like transitivity and density, and also connectivity analysis by computing and studying the properties of the connected components in the network.

2 We study how entities affiliated with different countries interact and manifest in the Panama Papers by doing a country analysis both at the network level, and at the connected component level. We use both a traditional metric like country assortativity in support of this analysis, as well as the information-theoretic *entropy* metric, which is much less commonly used in the network science community, to illustrate some key properties. For example, our entropy analysis shows that, while smaller components have fewer unique countries represented in them after

[1] A review of this impact is provided in the "Related work" section.

[2] We say 'information' rather than 'money' because the relationships in the networks we study often indicate either the presence of a transaction, or a social/professional connection, but not the official records of an actual transaction.

controlling for size (smaller entropy), the larger components are much more diverse, in addition to being much denser (higher entropy).

3 Using the basic networks constructed earlier, we construct a set of *higher-order networks* that illustrate novel phenomena at scale that may provide structural insights into corruption and other illicit activities that are emblematic of the Panama Papers. Using three such higher-order networks, we find some key unexpected deviations from what we would expect from such networks, especially significant violation of the power law distribution that is typically expected in the degree distributions of such networks. In contrast, no such deviation is noted in the connected component size distribution. While the causes of these deviations are likely complex, we believe that they may be suggestive of unusual (and potentially illicit) activity that is peculiar to the information leaked in the papers.

The rest of this article is structured as follows. We begin with a synthesis of related work, followed by a description of the data. We then describe the structural study of selectively constructed networks, followed by our proposed construction and analysis of higher-order homogeneously typed networks. We show that these networks illustrate key non-random deviations that may be fertile ground for new theory in the context of the activities known to be represented in the data. We conclude the paper by summarizing our main findings and providing a brief description of planned future research and opportunities.

## Related work

Since the release of the Panama Papers by the International Consortium of Investigative Journalists (ICIJ), multiple pieces of work, including at least one bestselling book, have sought to study the information in them and reveal how the 'rich and the powerful' hide their money. The book by Obermayer and Obermaier (2016) is an instrumental resource in this regard and describes the leak as much broader than the data that was eventually released to the public for analysis. The full dump eventually amounted to more than 11.5 million documents, delivered in real-time installments, and the book makes an early reference to important terms encountered throughout this article including 'offshore' entities and 'intermediaries' (which are described by the book to be 'lawyers, accountants and banks').

However, it is worthwhile noting that, despite the several colorful case studies (e.g., one of the early chapters is titled 'Vladimir Putin's mysterious friend' (Obermayer and Obermaier 2016)) and the release of some vital statistics and historical background (e.g., that the papers included the records of more than 214,000 offshore companies), the book does not attempt to provide a structural or scientific analysis of the entities. In contrast, in this article, we conduct just such an analysis rather than selecting and studying specific instances of illicit entities or actors. It should be noted that ICIJ itself claims on its website, dedicated to the Panama Papers (we provide a link in the next section), that mere presence of an entity in the Panama Papers does not suggest that it is engaged in shadowy or illicit activities. In this article, in keeping with this recommendation, we take a more agnostic view and aim to uncover structural insights using the framework of network science. In the context of the Panama Papers, to the best of our knowledge, this is the first such article to perform such a study.

Numerous other references have built off of (Obermayer and Obermaier 2016), including work by Trautman (2016) on disclosures regarding the leaked documents,[3] a discussion of the impact of bribery and corruption on the global community and issues surrounding tax evasion, (O'Donovan et al. 2019) on how firms use secret offshore vehicles to 'finance corruption, avoid taxes and expropriate shareholders,' and at least ten others[4] (Cooley et al. 2018; Ellis 2019; Heimann and Pieth 2017; Graves and Nabeelah 2019; Miller 2017; Miethe and Menkhoff 2017; Neu et al. 2020; Radon and Achuthan 2017; Tuttle 2016) and (Nerudova et al. 2020). Many of these works are more sociological, philosophical or legal in nature e.g., the work by Trautman was published in a law review, while the article by Radon and Achuthan (2017) was published in international affairs. Some of the books are much more specific in nature e.g., (Jancsics 2018) and (Rotberg and Carment 2018) respectively consider the issue from the lens of shell companies (and government corruption) and Canadian corruption (both foreign and domestic). Computational studies of any kind have not been common; one rare (but still relatively non-computational) example is the work by McGregor et al. (2017), which (although published in the USENIX security symposium that tends to cover computational work), uses survey data from participating journalists and other stakeholders to understand how a large and diverse group of globally dispersed journalists met and maintained key security goals (including secrecy of the project) until the launch date. In the work that is most similar to this one, but still limited to a geopolitical region and to relatively constrained structural analyses, is the effort by Rabab'Ah et al. (2016) to study the financial networks of the Middle East. For the ethical issues pertaining to use of datasets such as the Panama Papers for research, we refer the reader to the work by Thomas et al. that directly references the papers (Thomas et al. 2017). Another computational work that involves information extraction, but is markedly different from this article is the multilingual pipeline proposed by Wiedemann et al. (2018).

By using network science to provide a scientific view of the structure in the Panama Papers, we hope to add to this body of work and to allow experts in sociology and law to tie in their own expertise to some of the structural variables expressed in this paper, and to gain an even deeper understanding at scale of the papers.

Beyond the Panama Papers, network science has, of course, been used to study an increasing variety of complex systems in the recent past (Barabási and et al. 2016). The ubiquity of the power law-obeying degree distribution, in Web and other modern large-scale networks, is itself quite a recent finding, dating to less than 25 years ago. A non-comprehensive, but fairly representative, set of papers that have sought to study complex systems (some of which use Big Data) using applied network science include (Borgatti et al. 2009; Chen and Redner 2010; Hummon and Dereian 1989; Gavin et al. 2002; Li et al. 2007) and (Greenberg 2009).

## Data

The International Consortium of Investigative Journalists (ICIJ) has shared the Offshore Leaks Database by licensing it under the Open Database License and its contents under the Creative Commons Attribution-ShareAlike license. The data comes from Panama law

[3] From the Panama-based law firm Mossack Fonseca.

[4] We attempt to provide a representative list that includes works from across the spectrum, including social sciences, social media and philosophy.

firm Mossack Fonseca, whose inner workings were exposed in the journalistic investigation published in April 2016 in conjunction with Süddeutsche Zeitung and more than 100 media partners.We use the latest version of the dataset that is currently available on their project page.[5] There are three main kinds of entities that are relevant to the studies in this paper. We take their definitions from the ICIJ page for context:

1 *Offshore Entity:* An offshore entity is a company, trust or fund created by an agent in a low-tax jurisdiction that often attracts non-resident clients through preferential tax treatment. Herein, an agent is a firm in an offshore jurisdiction that incorporates, registers and manages an offshore entity at the request of a client.

2 *Intermediary:* An intermediary is a go-between for someone seeking an offshore corporation and an offshore service provider – usually a law-firm or a middleman that asks an offshore service provider to create an offshore firm for a client.

3 *Officer:* A broad class of individuals who are in a position of significant influence in the associated organization. Examples include a nominee (a person or company that acts on behalf of the beneficial owner of an entity to provide an extra level of secrecy), a beneficiary (a person who is entitled to certain financial benefits under a trust arrangement. Sometimes beneficiaries are not aware of their role in a trust because the neither the settlor nor the trustee has notified them) and a protector (adviser to a trust settlor who oversees the work of the trustee).

Within the dataset collection downloaded from the links above, several files are exposed. The most important of these is an edge list that expresses a *global multi-relational Panama network* (referred to succinctly as the Panama network unless otherwise noted), defined as the directed network of all *labeled* edges (called *triples*) or the set $\{(a, R, b)\}$ where $a$ and $b$ are members of the set comprising all offshore entities, intermediaries, officers and addresses in the dataset, and $(a, R, b)$ means that $a$ and $b$ are directly related using some relation $R$. The Panama network is global because the nodes span the globe, and it is multi-relational since there are 33 unique relations. Each relation falls under one of three relation types, namely 'registered_address', 'intermediary_of' and 'officer_of'. In total, there are 657,530 unique triples in the Panama network. Of the 33 unique relations, one each has type 'registered_address' and 'intermediary_of', while the other 31 are typed as 'officer_of' relations, examples including 'sole shareholder of' and 'power of attorney of'. Note that the term 'officer' here is defined loosely, and does not necessarily have to refer to an employee of the company i.e. any individual explicitly associated in an influential role with an organization is linked to that organization using a typed 'officer_of' link. In practice, we found that the specific relations 'shareholder of' and 'beneficiary of' dominated the 'officer_of' distribution, with the other 29 relations typed as 'officer_of' contributing an extremely small weight to the probability distribution.

Table 1 records some vital statistics, including node counts (and singleton and non-singleton counts of these nodes) by type. By *singleton*, we mean nodes that have degree 0 i.e. are not linked to any other nodes via an applicable relationship or edge. A non-singleton is any node that is not a singleton. Since we are studying the structural and relational properties of the Panama Papers in this work, singleton nodes do not contribute to the study; furthermore, their extremely small proportion compared to the non-singletons provides some assurance that they have limited empirical impact. It is

[5] https://offshoreleaks.icij.org/pages/database

**Table 1** Some node statistics on the global Panama multi-network

| Node Type | Singletons | Non-singletons | TOTAL |
|---|---|---|---|
| Address | 12 | 93,442 | 93,454 |
| Intermediary | 36 | 14,074 | 14,110 |
| Officer | 119 | 238,283 | 238,402 |
| Offshore Entity | 0 | 213,634 | 213,634 |

telling, however, that the vast majority of entities in the Panama Papers have some connection to some other entity. Thus, the Panama network is naturally amenable to the kind of structural analysis we undertake in this work. If most of the nodes had been singletons, instead of non-singletons, an argument could have been made that network science is not the correct scientific framework to be studying this complex system, since there would have been no apparent structure for a network to model (at least, without extra information that is not currently available in public datasets). Figure 1 also provides a probability distribution of the 33 relations in terms of the number of triples they occur in. When broken by type (not shown in the figure), there are 151,105 registered_address, 213,634 intermediary_of and 309,363 officer_of links.

### Data preprocessing and statistical profiling

Although Fig. 1 provides some aggregate statistics on relation frequencies, it is useful to study the Panama network as a whole using network-theoretic measures. To calculate ordinary network-theoretic measures, we first model the directed, labeled network as an *unlabeled* multi-network by removing edge labels, but by retaining the multiplicity of edges between two nodes to indicate that more than one relation exists between those nodes. We also consider the 'simple' version of this network, where multiplicity of edges

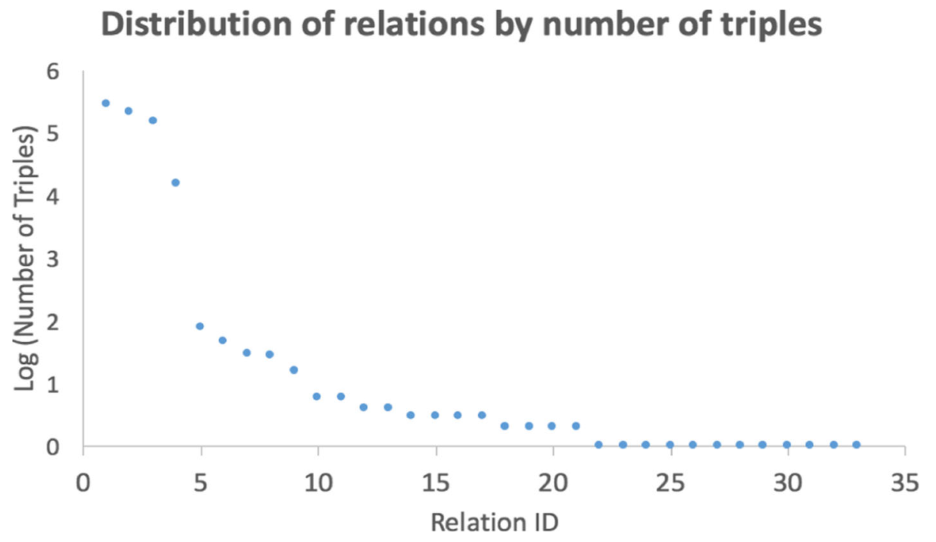

**Fig. 1** Probability distributions (on a semi-log scale) of the 33 unique relations in the dataset, with IDs assigned to relations according to descending order of frequency (number of triples containing that relation)

**Table 2** Network measures on the Panama multi-network. Density, and other such metrics, were computed with respect to non-singleton nodes e.g., the density of the simple graph underlying the Panama multi-network is 657,489/(559,433*559,432/2) which is approximately 4.202e-06 (as reported below). Even among the non-singleton portion of the network, there are no non-singleton strongly connected components; hence, the total number of strongly connected components is simply equal to the number of non-singleton nodes (i.e. the number of nodes in the largest strongly connected component is simply 1)

| **Network Measure** | **Value** |
|---|---|
| Number of Non-Singleton Nodes | 559,433 |
| Number of Edges in Simple Graph | 657,489 |
| Simple Graph Transitivity | 5.112e-08 |
| Simple Graph Density | 4.202e-06 |
| Degree Assortativity Coefficient (only applicable to multi-graph) | -0.051 |
| Number of Weakly Connected Components (WCCs) | 11,043 |
| Number of Nodes in Largest WCC | 455,479 |
| Number of Strongly Connected Components (SCCs) | 559,433 |
| Number of Nodes in Largest SCC | 1 |

is not retained.[6] A snapshot of the network-theoretic statistics is provided in Table 2 and the in-degree and out-degree distributions of the multi-network are illustrated in Fig. 2.

Figure 2 seems to suggest that the distributions obey a power-law (like many previously observed degree distributions in information and social networks) but estimating the power-law exponent[7] (also called the *scaling parameter*) using least-squares linear regression on the logarithms of the degrees and their empirical frequency can be methodologically problematic and likely to lead to sub-optimal results (Clauset et al. 2009). The main reason is that, for most empirical phenomena, the power-law tends to apply only for values greater than some minimum value $x_{min}$. For a particular distribution, given $x_{min}$, and empirical observations of a discrete variable $x$ (which is true in our case, since $x$ represents node degrees), we can use Eq. (3.7) from (Clauset et al. 2009), which draws on an approximate form of a non-analytic expression obtained using Maximum Likelihood Estimation (MLE):

$$\hat{\alpha} \approx 1 + n \left[ \sum_{i=1}^{n} ln \frac{x_i}{x_{min} - 0.5} \right]^{-1} \quad (1)$$

Here, $x_i$ is an element of the set of all $x$ such that $x \geq x_{min}$. We visually estimated $x_{min}$ for the in-degree and out-degree distributions to be 5.96 and 2.8, respectively. Applying this value and the empirical degree observations in Eq. 1, we obtain the respective exponents as 3.315 and 2.186. Interestingly, while the power-law exponent of the out-degree distribution conforms to expectations (in most networks, the exponent is found to range from 2 to 3), the in-degree distribution has a steeper slope on the doubly-logarithmic plot.

The *density* of a network (whether simple or multi-graph) is the number of edges divided by the total possible number of edges in the simple equivalent network. In theory, this allows the multi-graph density to exceed 1, though we do not observe this possibility in our sparse network. The *degree assortativity* is the Pearson correlation coefficient

[6] Once edge labels are removed, a set of unique triples does not have to necessarily yield a set of unique *unlabeled edges*. For example, Officer A might be vice president of Offshore Entity B, and A might also be a shareholder of B. In the labeled network, these are two unique triples (relations). In the unlabeled multi-network, there are two edges between A and B, while in the unlabeled simple network, there is only one edge.

[7] The parameter, $\alpha$, when $p(x) \propto x^{-\alpha}$.

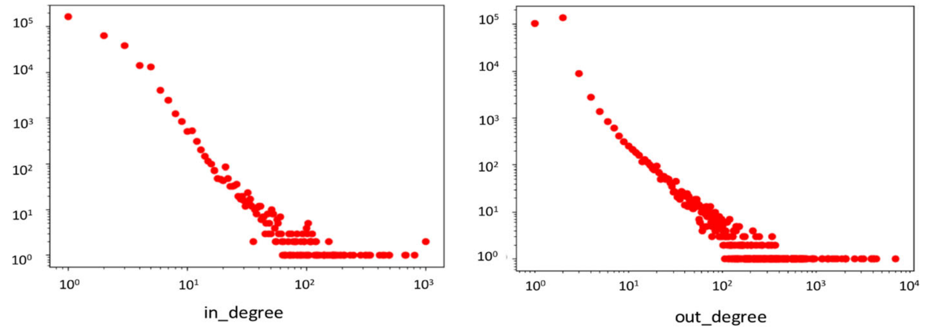

**Fig. 2** In-degree and out-degree distributions of the unlabeled, directed Panama multi-network. Using a method first proposed in Clauset et al. (2009) (briefly summarized in the main text), we estimated the in-degree and out-degree power-law coefficients to be 3.315 and 2.186 respectively

of degree between pairs of linked nodes. A *weakly connected component (WCC)* is a connected component where a path exists between every pair of nodes in the component regardless of edge directionality. However, in a *strongly connected component (SCC)*, every pair of nodes must be linked via a path with directed edges i.e. directionality of edges cannot be ignored when computing the path. The table shows that there are no non-trivial SCCs in $G$ (every node falls in its own SCC when partitioning the graph into SCCs). *Transitivity* has a similarly established definition: it is the fraction of all possible triangles present in the graph. Transitivity for multi-graphs is not well-defined; hence, we only show it for the simple network equivalent of the Panama network.

Overall, the network is extremely sparse and unlike social networks, the transitivity is also very low. Although not shown in the table, the network was also found to be 'almost' simple; the difference between the simple and multi-network was found to be miniscule.[8]

Figure 3 also illustrates a triad census of the simple network. Triads are both theoretically and practically important in complex networks, as several lines of evidence have illustrated over the decades (Antal et al. 2005; Moody 1998; Wasserman 1977), . For a recent reference, we refer the reader to (Faust 2010). The nomenclature for these triads (e.g., 003 indicating the null triad that has no links between the three actors) is fairly standard in the literature. The triad census in Fig. 3 shows that the triad with no links (003) dominates by orders of magnitudes. The network is not symmetric by nature, as the difference between the triad counts for 012 and 102 illustrate. While there are billions of triads with one uni-directional edge, there are no triads with one bi-directional edge. There are a reasonable number of triads (021U and 021D) with two edges, though a triad such as 021C is still quite rare.

We conducted statistical significance analyses on all triad counts by first using *degree-preserved edge shuffling*[9] that is often used in the literature to generate random graph topologies with the same size and degree as the 'target' graph (Elhesha and Kahveci 2016; Gale and et al 1957; Milo et al. 2003). Specifically, we use this technique to generate 100 random graphs, of which both the in-degree and out-degree distributions are preserved compared to $G$. We computed the z-statistic of all motifs in Fig. 3, and found that the

[8] Specifically, the multi-graph density was 4.308e-06 and the number of edges in the multi-graph was 657,530. Both measures are very similar to that of the simple graph shown in Table 2.

[9] Briefly, we do the following operation $10 \times |E|$ times: we randomly remove two edges $(u_1, u_2)$ and $(v_1, v_2)$ from $G$, swap $u_1$ and $v_1$, and insert the 'new' edges $(v_1, u_2)$ and $(u_1, v_2)$ into $G$.

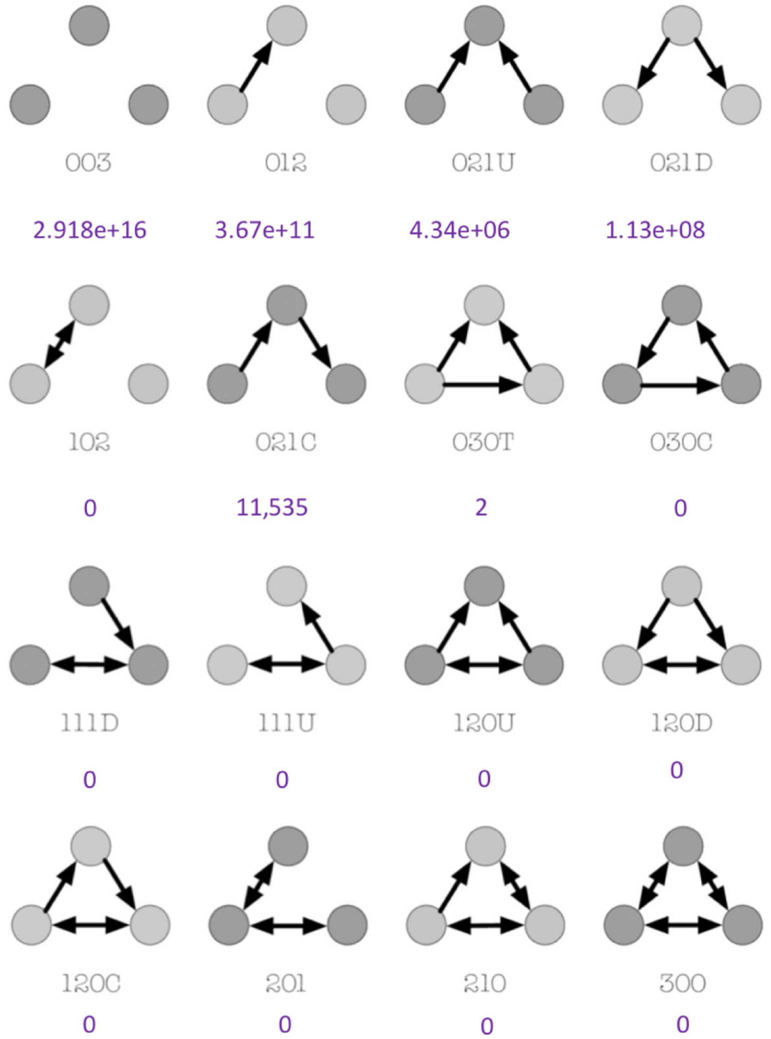

**Fig. 3** A triad census of the simple network equivalent of the Panama multi-network

absolute value of the z-statistic was much higher than 3 for triads 003, 012, 021D, 021U, 021C and 030T.[10] Of these, the z-statistic for triads 021U, 021C and 030T were all negative, attesting to the fact that this network is much sparser in triangles (030T) and triples (021C and 021U) than even random graphs with a similar degree profile.

The triad census seems to suggest transactions of a one-way nature, or one that favors secrecy, since three-way interactions are statistically dampened compared to random graphs with similar topology. This is the first indication that the Panama Papers, though a complex system, may not obey the usual laws that other complex systems with social actors and relationships seem to. Later, we provide more evidence that lends further credence to this claim.

According to Table 2, the network also has small negative degree assortativity, indicating some measure of preferential attachment. In part, some of these numbers may be due to the heterogeneous nature of the graph, since the presence of 'address' nodes skews

[10] Triads 102, 111D, 111U, 030C, 201, 120D, 120U, 120C, 210 and 300 had count of 0 in all random graphs as well; hence, their 0 counts in $G$ are not significantly different from that of the random graph. The z-statistic was undefined, since it equated to 0/0.

some of the results. The reason is that, as shown in Fig. 4, address nodes do not have any outgoing edges, and a similar skew is noted for officer and intermediary nodes (no incoming edges). In the next section, we consider derivatives of this directed unlabeled multi-network that allow us to compare the network structure without this skew, and in ways that are more interpretable.

## Structural study of selectively constructed networks

Although the metrics computed on the global Panama network are instructive, they have limited interpretive validity, due to considerable heterogeneity both in the types of nodes and relations used in the network model. To obtain useful insights, we derive three other directed, unlabeled networks from the original set of triples. We also consider, by way of a baseline, the unlabeled multi-network in the previous section as a fourth unlabeled, directed multi-network designated as $G$. These three other networks have specific link semantics, as we describe below, and hence, are more amenable to tractable and interpretable structural analysis.

1. $G_o$: This network is constructed by only retaining links in $G$ that are typed as *officer_of* in the full graph $G$.
2. $G_i$: This network is constructed by only retaining links in $G$ that are typed as *intermediary_of* in the full graph $G$.
3. $G_{io}$: This network is the union of the two networks above. It is also the same as $G$ except with *registered_address* links removed.

One caveat from the description above is that, while we do control for links in the networks below, the nodes are still heterogeneously typed i.e. the *officer_of* network ($G_o$) contains both officers and non-officers, since the *end_id* of a triple that has an *officer_of* type relation is a non-officer. In a later section, we also consider 'higher-order' networks where we perform a systematic closure to obtain networks with homogeneously typed nodes.

Once constructed, we do not distinguish between nodes of different types; nor are link labels considered. Hence, the network becomes unlabeled (at the edge level) and untyped (at the node level). This is necessary for computing standard structural metrics. Specifically, we study and compare these networks from various perspectives (using $G$ as a baseline, where appropriate), first by plotting the degree distributions in Fig. 5. We find

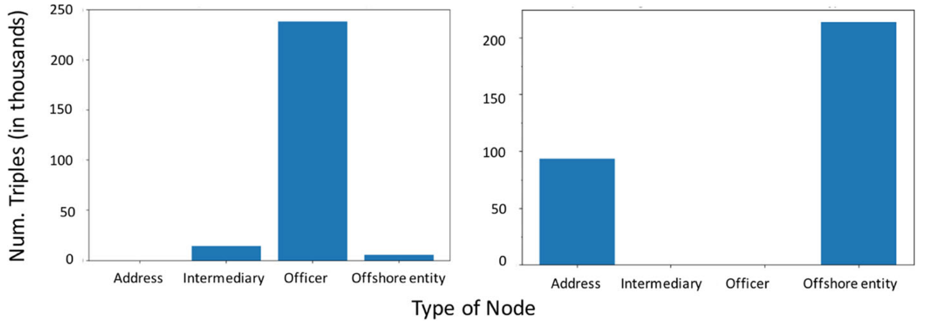

**Fig. 4** Number of triples (each triple being of the form (start-id, relation, end-id)) where a node of the type shown on the x-axis occurs as the start-id node (left) and end-id node (right)

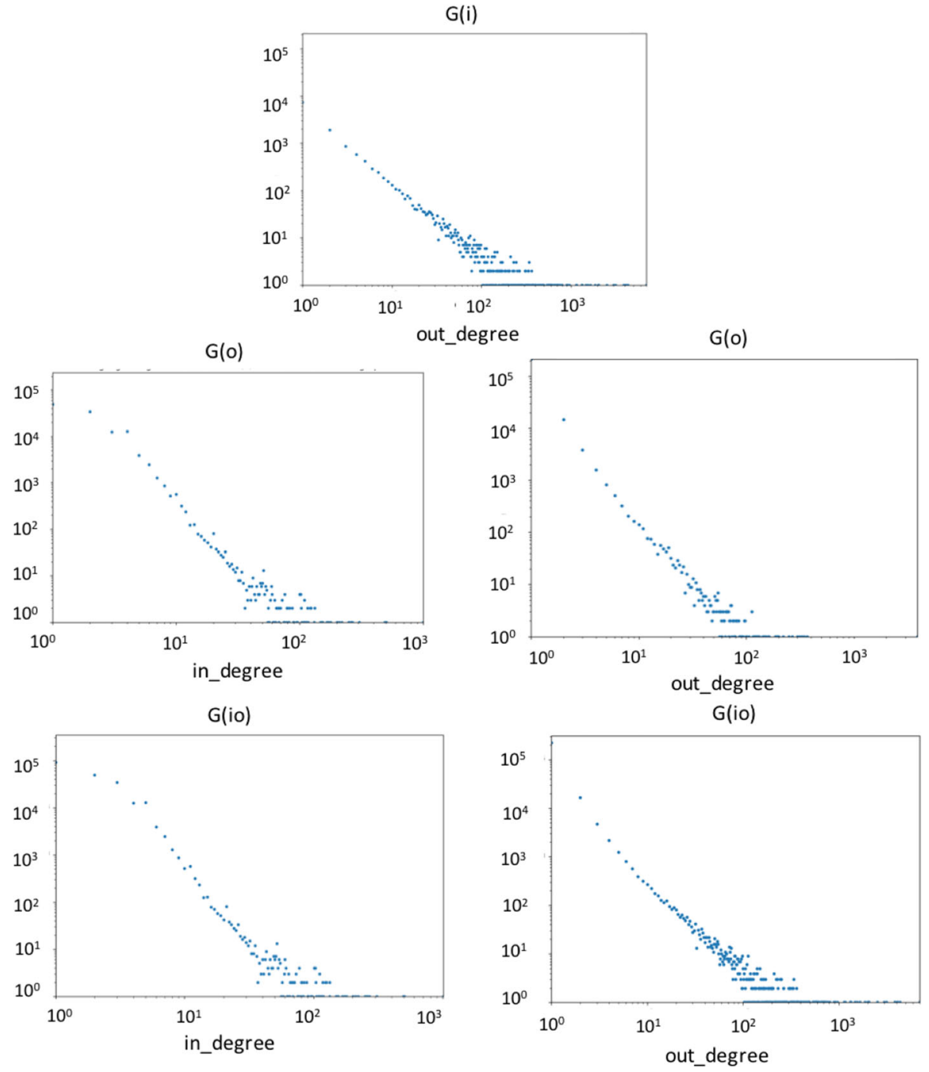

**Fig. 5** In-degree and out-degree distributions (on a log-log scale) of $G_i$, $G_o$ and $G_{io}$. $G_i$ has a trivial in-degree distribution as nodes have either degree 1 or 0 with frequencies 213,634 and 14,074 respectively. For this reason, we do not reproduce it as a plot below. The y-axis represents the empirical frequency of the degree in all cases

once again that the distributions are power-law, with only one exception (the in-degree distribution of $G_i$). At first glance, this would seem to suggest that these networks are not very different from ordinary social or organizational networks, which tend to have similar power law distributions. However, some other metrics, subsequently described, paint a different picture.

To assess connectivity, we consider the undirected equivalents of these networks, and calculate the distribution of connected components as a function of size (in terms of number of nodes within the connected component). The connected component size distributions also obey a power-law distribution (Fig. 6), showing that, although the networks are disconnected, there is high connectivity in large portions of the networks. Lack of connectivity, and a systematic distribution of component size, is one indication that the

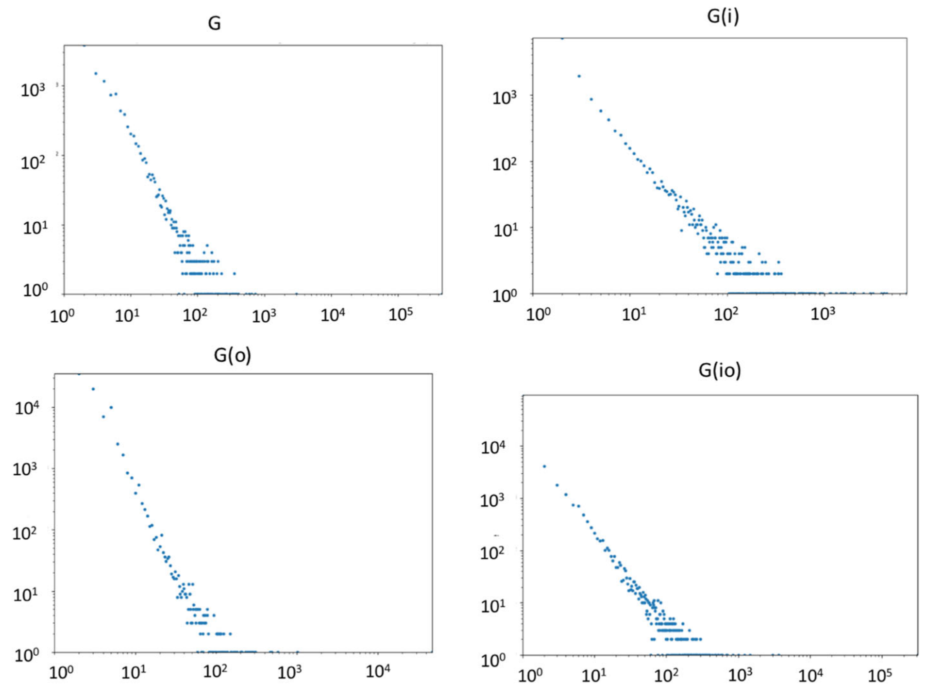

**Fig. 6** Connected component size distributions of $G$, $G_i$, $G_O$, and $G_{iO}$. The x-axis is the number of nodes in a connected component and the y-axis is the frequency of that size in the dataset

networks are *dissimilar* from social networks which tend to be connected (or almost connected), and where the diameter of the network has been shown to be decreasing over time due to densification phenomena (Leskovec et al. 2007). In other words, the Panama Papers do not seem be exhibiting any 'small-world' phenomenon to the extent supported by available data (Watts 2004).

The specific power-law exponents are tabulated[11] in Table 3. Not including the in-degree distribution of $G_i$, of which the exponent is non-interpretable due to the trivial nature of the distribution, the in-degree distribution of the other networks exhibit a higher exponent than the out-degree distributions. While all the values are between the range of 1.5 and 3.5, the usual range in ordinary social and information networks is between 2 and 3, which is only true for the in-degree distribution of $G_o$ and the out-degree distribution of $G$ in the table. The connected component size distribution exponent is close to 1, except $G_o$, where it achieves a high value of 1.568.

Many of the network measures computed in Table 2 are also computed for the selectively constructed networks in Table 4. In comparing the statistics in Table 4 with those of the original network in Table 2, we observe that *officer_of* nodes are more heavily represented in the non-singleton portion of the network than intermediaries (almost  359k vs. 228k), an increase of 57%, whereas the increase in the number of edges is relatively smaller (37% increase for the simple network and 45% increase for the multi-network). Transitivity in all networks is very low, either zero or near zero. Density is very low in all networks as well, and the assortativity coefficient (small negative values throughout) shows consistent but small disassortativity. Similar to the original network (though not tabulated here),

[11] The approximation method (Clauset et al. 2009) that was proposed earlier in the paper for estimating the power-law exponents is used again.

**Table 3** Exponents for the various power-law distributions characterizing the structure of $G$, $G_i$, $G_o$, and $G_{io}$. Because of the trivial nature of the $G_i$ in-degree graph, its in-degree exponent does not have the usual interpretive validity (as it does not remotely resemble a power law distribution); hence, we denote it with an x

| **Power-law Exponent** | **G** | $\mathbf{G}_i$ | $\mathbf{G}_o$ | $\mathbf{G}_{io}$ |
|---|---|---|---|---|
| In-degree distribution | 3.315 | x | 2.595 | 3.491 |
| Out-degree distribution | 2.186 | 1.506 | 1.807 | 1.689 |
| Connected component size distribution | 1.118 | 0.998 | 1.568 | 1.186 |

all networks were once again found to be ‘almost’ simple. The number of edges in the multi-network equivalents of $G$, $G_i$, $G_o$ and $G_{io}$ were respectively found to be 4.308e-06, 8.240e-06, 4.803e-06 and 3.342e-06 (nearly identical to simply graph density). A similar observation was made for the number of edges in the multi-network equivalents.

We also compute and tabulate the local and ‘meso-scale’ metrics of the four networks in Table 5. As a first step, we tabulate the *maximum core number* of both the whole network and the largest component, which turned out to be the same for all four networks. To compute the maximum core number for a graph, we note first that a *k-core* is a maximal subgraph that contains nodes of degree $k$ or more. The *core number* of a *node* is the largest value $k$ of a k-core containing that node. To compute the maximum core number of a graph or sub-graph, we simply compute the maximum core number observed in that graph, with the maximum taken over all nodes. The maximum core numbers show that there are ‘tight clusters’ where all nodes have degree 6 or 7. $G_i$ is an exception, which shows that the intermediary network behaves differently from the other networks in this regard.

A *bridge* in a connected component is an edge that, were it to be removed, would lead in the connected component ‘breaking’ into two connected components. The results in Table 5 show that the percentage of bridges (as a percentage of the number of edges) is high in all networks, and is extreme (100%) in the largest component in $G_i$. This suggests that every edge in the largest component in $G_i$ is a bridge, which is true for graphs that are star-like or linear chains (or a combination of the two).

We also studied the core-periphery structure of each network by using the Kojaku-Masuda algorithm with the configuration model[12] (Kojaku et al. 2019). This algorithm partitions the nodes in the component into a set of cores and a set of peripheries. For full details, including theoretical justifications on core-periphery models, we refer the reader to the seminal references in Baldwin et al. (2011); Forslid and Ottaviano (2003); Krugman (1991). Herein, we note that, the intuitive idea is to detect nodes where the network seems to be *concentrated* (the cores) with the peripheries representing the ‘outlying’ nodes. While a network is not inherently spatial (which is the context in which the modern core-periphery model was introduced, at least in Krugman’s work on New Economic Geography (Krugman 1991)), there is some intuitive semblance of concentration in most networks. In counting the number of cores in all networks, we see that only $G_i$ seems to have a single core, lending further credence to the earlier hypothesis that the network is star-like. The other networks have roughly equal numbers of cores and peripheries. This suggests that (i) either the algorithm is not appropriate for this kind of network, and a novel approach may be required to determine what the cores and peripheries are (if

[12] Implemented in the *cpalgorithm* package in networkX.

**Table 4** Network measures computed for $G$ (reproduced as a baseline from Fig. 2), $G_i$, $G_o$, $G_{io}$

| **Network Measure** | $\mathbf{G}$ | $\mathbf{G}_i$ | $\mathbf{G}_o$ | $\mathbf{G}_{io}$ |
|---|---|---|---|---|
| Number of Non-Singleton Nodes | 559,433 | 227,708 | 358,918 | 465,788 |
| Number of Edges in Simple Graph | 657,488 | 213,634 | 292,749 | 506,383 |
| Simple Graph Transitivity | 5.112e-08 | 0 | 0 | 0 |
| Simple Graph Density | 4.202e-06 | 8.240e-06 | 4.545e-06 | 3.236e-06 |
| Degree Assortativity Coefficient (only applicable to multi-graph) | -0.051 | -0.152 | -0.011 | -0.064 |

it is theoretically unjustifiable to have so many cores in the network), or (ii) there truly are many cores in the networks (which we claim is plausible, due to the transactional, global but highly decentralized nature of the entities in the Panama Papers[13]), suggesting a degree of robustness that we comment on in a later section. Since there are no clear cores, the problem of disrupting such a network becomes much harder for federal and transnational agencies tasked with minimizing the impacts of money laundering and organized crime.

We also computed the number of *undirected connected triples* in each graph.[14] A connected triple in an undirected graph is every 3-tuple of nodes $(a, b, c)$ such that there is an edge between $a$ and $b$ and an edge between $b$ and $c$. For Since the connected triple is undirected, triples $(a, b, c)$ and $(c, b, a)$ are treated the same (and only counted once). For example, a triangle would contribute three connected triples. On this measure, the difference emerges in $G_o$, which has far fewer connected triples (almost an order of magnitude less) than the other graphs. We believe that the reason lies in the network construction itself; namely if $A$ is an officer of $B$, then by definition, $A$ is an officer-type node, and $B$ is either an intermediary or an offshore entity. However, since $G_o$ only contains officer-type relations, the only way that $B$ could participate in a connected triple is (i) if $A$ is also the officer of *other* organizations, in which case $B$ would be the third (or equivalently, first) element of all connected triples where one of the organizations is at the 'opposite' end of the triple (either third or first element) with $A$ in the middle, (ii) if $B$ has other officers, in which case $B$ would be the middle element of (one or more) connected triples with the two officers (one of whom is $A$ serving as the 'ends' of the triples. Since both of these possibilities are likely, we do not observe a value of 0 for $G_o$; however, the observed value is still very low compared to other networks. This may, in turn, suggest that organizations in the Panama Papers simply do not share many officers, or that organizations do not have many officers to begin with. Since many companies in the Panama Papers are suspected to be shell companies, rather than real businesses, this result can be interpreted in the sociological context of the papers.

The raw s-metric of a graph is defined as the sum (over all edges $(u, v)$) of the quantity $d(u) * d(v)$ where $d$ is the degree function. The correct way to interpret this metric is across the four networks. Once again, we find that $G_o$ has smaller s-metric than the other networks. In general, this means that degrees of nodes in participating edges are simply not high compared to the other networks. However because the s-metric grows

[13] A further argument for this claim is that the algorithm was run over the whole network, which we know is very disconnected from the earlier analysis of the connected components.

[14] For directed graphs, a triad census conveys the structural properties of sets of three nodes in far more detail. The triad census for $G$ was illustrated earlier. We do not explore the triad census for the other networks as the metrics in Table 5 conveys the key small-scale and meso-scale properties of these networks more succinctly.

**Table 5** Local and 'meso-scale' metrics characterizing the structure of the *undirected* equivalents of $G$, $G_i$, $G_o$, and $G_{io}$. Details on the metrics, as well as the core-periphery algorithm used for computing the partition of nodes in each network into cores and peripheries are provided in the text

| Structural Metric | **G** | **G**$_i$ | **G**$_o$ | **G**$_{io}$ |
|---|---|---|---|---|
| Max. core number in largest component | 7 | 1 | 6 | 7 |
| Max. core number in network | 7 | 1 | 6 | 7 |
| Num. bridges in largest component (% of edges in largest component) | 306,025<br>(54.84%) | 7,016<br>(100%) | 35,337<br>(60.9%) | 257,527<br>(71.65%) |
| Number of core nodes | 233,337 | 1 | 23,643 | 163,204 |
| Number of periphery nodes | 222,142 | 7,016 | 23,803 | 147,471 |
| Number of connected triples (in millions) | 117.36 | 104.29 | 10.18 | 114.76 |
| Raw s-metric | 5.597e+8 | 2.08e+8 | 4.63e+7 | 5.43e+8 |
| Number of unique 3-clique motifs (triangles) | 2 | 0 | 0 | 0 |
| Number of unique 4-clique motifs | 0 | 0 | 0 | 0 |

very quickly with even a few reasonably high-degree nodes, the difference may not seem as stark as the numbers may suggest.

Finally, in considering the numbers of unique 3-clique and 4-clique motifs in the networks, we find that there are only 2 triangle motifs (or 3-cliques) in the overall network, corresponding to the two instances of triad 030T in Fig. 3, and no triangles in the other networks. This is in conformance with the extremely low transitivities and densities observed earlier for these networks. The number of 4-motifs was found to be 0 in all networks, which means that the maximum sized clique in all graphs (except $G$ where it is 3) is 2. The graphs are sparse, although they do have interesting structural properties. Taken together, our findings strengthen the claim that, while the Panama Papers embody a complex system with many actors and players spread across the globe, they do not seem to follow the same kinds of laws that other complex systems with social players seem to follow (such as friendship and follower networks of social media platforms).

### Country assortativity analysis

A key value proposition in conducting a study of this nature is to analyze the *country dependencies* in the various networks. We conduct such an analysis using two mechanisms. First, we compute *country assortativity* for all four networks (Table 6) after removing nodes that (i) either have no country associated with the node, or (ii) have more than one country associated with the node. Country assortativity is a special instance of the broader notion of *attribute assortativity* (where the attribute is the *country*, following the two pre-processing steps above), which is defined as:

$$r = \frac{tr(M) - ||M^2||}{1 - ||M^2||} \tag{2}$$

**Table 6** Country attribute assortativity for $G$, $G_i$, $G_o$, and $G_{io}$. Nodes in these networks that were associated with more than one country (or with no country at all) were not included in the analysis

| Network | Country Assortativity |
|---|---|
| $G$ | 0.699 |
| $G_i$ | 0.895 |
| $G_o$ | 0.314 |
| $G_{io}$ | 0.576 |

Here, $r$ is the assortativity coefficient, $tr$ is the trace and $M$ is the *mixing matrix* or the joint probability distribution of the specified attribute. Full details and analysis of attribute assortativity may be found in the seminar paper by Newman (2003). The intuition is that, if two nodes affiliated with the same country are linked in the network, the country assortativity would be positive and high; otherwise, it would be low (or even negative).

While nodes that have no country associated with the node, or that have more than one country associated with the node, may potentially be of interest to investigators (or are interesting elements of study in their own right), they constitute a relatively small fraction of the overall network; for each of $G, G_i, G_o$ and $G_{io}$, the percentage of such nodes (i.e. associated with no country or multiple countries) was found to be 17.4%, 1.2%, 26.2% and 17.4% respectively. The officer network is especially interesting because it seems to suggest that we either do not know the country associated with an officer, or the officer is associated with multiple countries, which may be a noisy artifact of the data. After removing such nodes, we are guaranteed to only have one country associated with each node, leading to a well-defined and interpretable analysis of country assortativity in the various networks.

We find that the country assortativity is generally high across all four networks, but is especially high in the intermediary network $G_i$. In other words, when A is an intermediary of B, they are very likely to belong to the same, rather than different, country. This is in accordance with what we would ordinarily expect, rather than in highly illicit networks. However, one caveat that should be noted is that the companies and intermediaries in the network are not 'ordinary' in the usual sense i.e. many of the organizations in the Panama Papers are 'offshore entities' that may themselves be associated with a bigger company or individual. Offshore entities in a given country would, for obvious reasons, prefer to partner with an intermediary in that country (i.e. a Swiss offshore entity would intuitively prefer to work with a Swiss law firm or accountant). In fact, the lower values of country assortativity for the other networks may suggest that offshore entities are set up in a country precisely to transact with intermediaries in that country.

Since the networks studied thus far have not been connected, we also conducted experiments via a second mechanism to understand how the country distributions are reflected in connected components of various sizes. The data in Table 6 seems to be suggesting that countries generally tend to co-occur together. However, this analysis is inherently limited because it only considers information at the edge-level. To gain a better sense of country representations and mixtures, we computed the *information-theoretic entropy* of the empirical probability distribution of countries within each connected component, again ignoring nodes that have no country or multiple countries. Specifically, the information-theoretic entropy may be defined as follows. Given an empirical probability distribution $P$, the information-theoretic entropy is given by $-P(x)\Sigma log_b(P(x))$ where the sum is taken over all discrete observations $x$. The base $b$ could be the natural exponent $e$, 2 or 10, but these do not give comparable or normalized values when there are many 0 entries, as is the case when computing the country entropy for a given connected component (i.e. most countries do not occur in the component and have probability 0). Therefore, we set $b$ to be the number of unique countries observed across all nodes in the connected component. Since this number can vary across components, $b$ can also vary with the component. An advantage of using a varying $b$, however, is that the entropy is always between 0 and

1, where a value skewing towards 0 indicates that very few countries (and in the case of 0, only one country) are represented in the component, while a value that skews towards 1 implies high diversity. As a result, the entropies across components become comparable, as they are on a normalized [0.1] scale.

A plot of entropy vs. connected component size (in terms of number of nodes in the component) is reproduced in Fig. 7. We find, intriguingly, that while there is a negative relationship between entropy and component size (implying that, on average, smaller components tend to be much more diverse while still retaining connectivity, though many small components also have entropy 0 or close to 0), the very largest component (in $G$, $G_o$ and $G_{io}$) has an entropy that is reasonably high (in the range of 0.6), which is unusual considering the size of the component. It suggests coordinated activity across companies and intermediaries in different (but not too many) countries, and may only be amenable to investigation (for illegal or nebulous activity) if national agencies agree to cooperate. By way of contrast, when we plot the *density* of the network instead of entropy (Fig. 8), components with more nodes are found to be less dense, exactly as would be expected for ordinary networks. The density of the largest component is only slightly above 0, suggesting high sparsity, even though the component (by definition) comprises a set of nodes that are connected to one another via at least one path. The difference between the sizes of the largest and second-largest components is also significant in all networks except $G_i$. This suggests that the largest component may be an interesting subject of study in its own right.

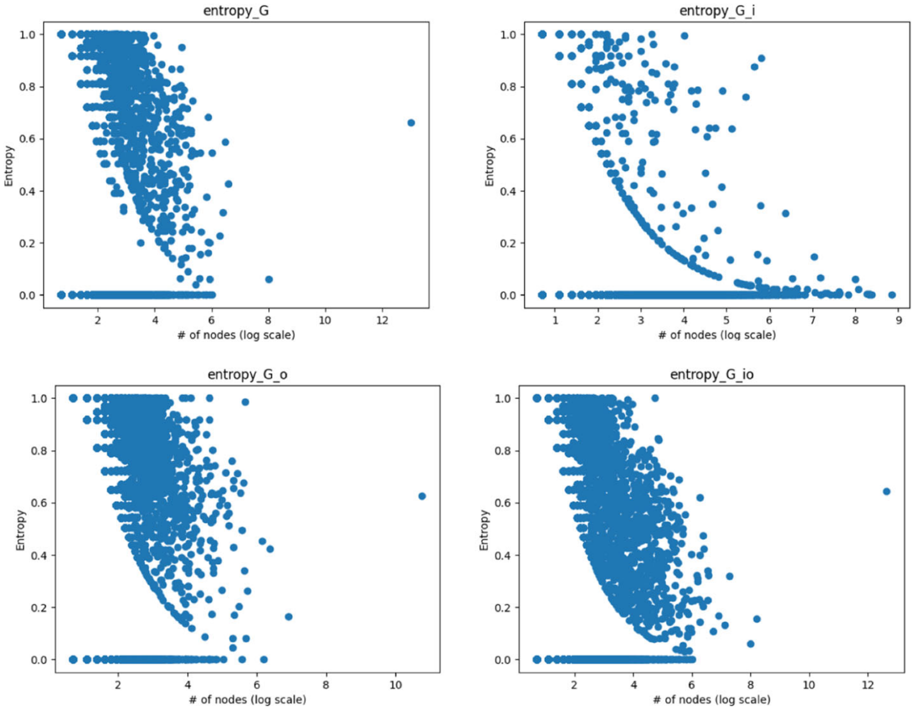

**Fig. 7** Entropy of connected components in $G$, $G_i$, $G_o$, and $G_{io}$ vs. size of connected component measured using number of nodes. The text provides an explanation of the base $b$ used in the logarithm of the entropy formula. The plots are on a semi-log scale, where the x-axis is on the log scale and the y-scale is on the ordinary scale in the [0,1] range

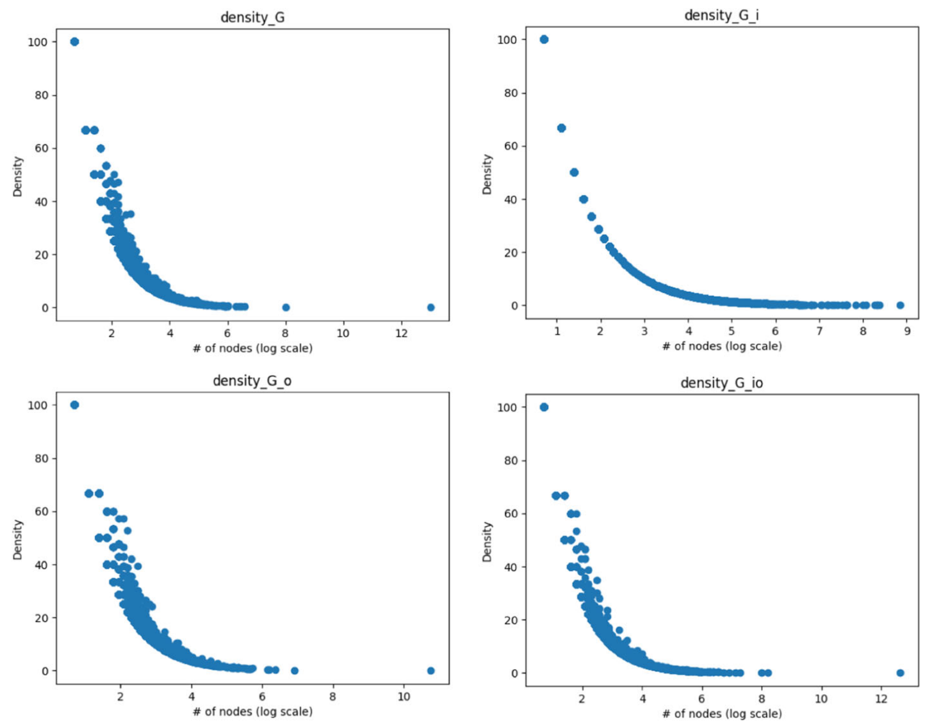

**Fig. 8** Density (expressed as a percentage) of connected components in $G$, $G_i$, $G_O$, and $G_{io}$ vs. size of connected component measured using number of nodes. The plots are on a semi-log scale (where the x-axis is on the log scale and the y-scale is on the ordinary percentage scale in the [0,100] range)

### Higher-order homogeneously typed networks

An issue with the networks constructed thus far is that they do not model (and in fact, completely ignore) potential relationships that are *implied* between nodes of the same type *through* nodes of different types. For example, two individuals may be officers at the same organization, but are not directly linked in the network (in fact, there is no relationship in the Panama Papers data, as currently available, that models any connection between two officer-type nodes). To model a connection between these officers, we have to take into account the fact that they share a neighbor (in the overall graph) that belongs to an organization-type node (i.e. the node is either an offshore entity or intermediary). By adding an edge between two nodes (of the same type) in the 'first-order network' (of which we saw four examples in the previous sections) if they indeed share such a (differently typed) neighbor, we obtain a more complete network that we denote as a *higher-order homogeneously typed network*. We consider three such networks in this section:

1 $G_o^*$: This is a network of officers, where we declare a link between two officers A and B iff (i) there is a direct link between them in the original triples set, (ii) if there are two triples in the triples-set of the form $(A, officer\_of, C)$, $(B, officer\_of, C)$ i.e. the two officers share a common organization. The second type of link (which is 'indirect') constitutes almost all links in the network as there are virtually no direct links between officers. We believe that this has preempted a structural study of an 'officer' social network implied by the Panama Papers.

2 $G_o^o$: The nodes in this network are offshore entity nodes, and an edge exists between two offshore entities iff (i) there is a direct link between the offshore entities (which

is never the case empirically), (ii) the two offshore organizations share an officer.

3 $G_o^i$: This network is similar to the above (nodes are offshore entity nodes) but we create a link between two offshore entity nodes if they share an intermediary.

Basic network statistics are reproduced in Table 7. Similar to what we observed for the first-order networks, the higher-order networks were also found to be 'almost' simple. The edges and density of the multi-networks were both found to be nearly identical to those of the underlying simple networks, as noted in Table 7. A major difference that we start to observe, at least for $G_o^*$, is higher assortativity and transitivity. This implies that $G_o^*$ resembles a social network much more than the first-order networks did. While densities are still low, they are still higher by an order of magnitude compared to the densities tabulated earlier in Table 4.

Most important, unlike the earlier networks, the degree distributions of the higher-order networks in Fig. 9 are unusual in that they do not obey a power-law distribution. There is a significant deviation, in particular, in the mid-degree range; in the $G_o^*$ network, a set of nodes with degree 100 or higher exhibit deviation from the power law and have higher frequency than the law would indicate. These nodes may be worth investigating from a structural and country standpoint, though we do not conduct this analysis within the work. There is a definite curvature in the $G_o^o$ network, but towards the right of the curve, we see a spread (across frequencies) of nodes with very high degrees. Perhaps the most interesting curve with respect to degree distribution is $G_o^i$, which initially seems to express the usual power-law distribution, but then reverses and exhibits another power-law distribution with a *positive* exponent, a rare occurrence that we could not find an explanation or theory for in the existing network science literature.

One hypothesis (for explaining these unusual observations) that we are currently investigating through further analysis is that the offshore entities in Fig. 9c can be sub-divided into different *markets*. Smaller offshore-entities may be set up, for example, to fulfill a formal requirement in the country in which they hope to transact or do business with another entity (possibly through an intermediary). Larger offshore-entities may be set up for acquisitions or other purposes. If we are able to systematically segment offshore entities into such 'markets', we may be able to observe a standard power-law distribution for some of the markets, but others may yield interesting insights obeying a separate set of laws.

A similar theory may apply to the other higher-order networks. For example, in both of the other higher-order networks illustrated in Fig. 9 there is deviation in some sections of the networks, but overall, the networks seem to follow an approximate power-law distribution. The deviation may also be suggestive of a separate 'market' (e.g., there are different

**Table 7** Network measures computed for the higher-order networks ($G_o^*$, $G_o^o$, and $G_o^i$), described in the text. NA means that the metric could not be computed for that network due to computational complexity or resource limitation issues

| **Network Measure** | $\mathbf{G_o^*}$ | $\mathbf{G_o^o}$ | $\mathbf{G_o^i}$ |
|---|---|---|---|
| Number of Non-Singleton Nodes | 200,682 | 45,599 | 206,332 |
| Number of Edges in Simple Graph | 1,638,339 | 8,332,183 | 104,290,798 |
| Simple Graph Transitivity | 0.9707 | NA | NA |
| Simple Graph Density | 5.771e-05 | 3.65e0-4 | 4.899e-03 |
| Degree Assortativity Coefficient (only applicable to multi-graph) | 0.7363 | NA | NA |

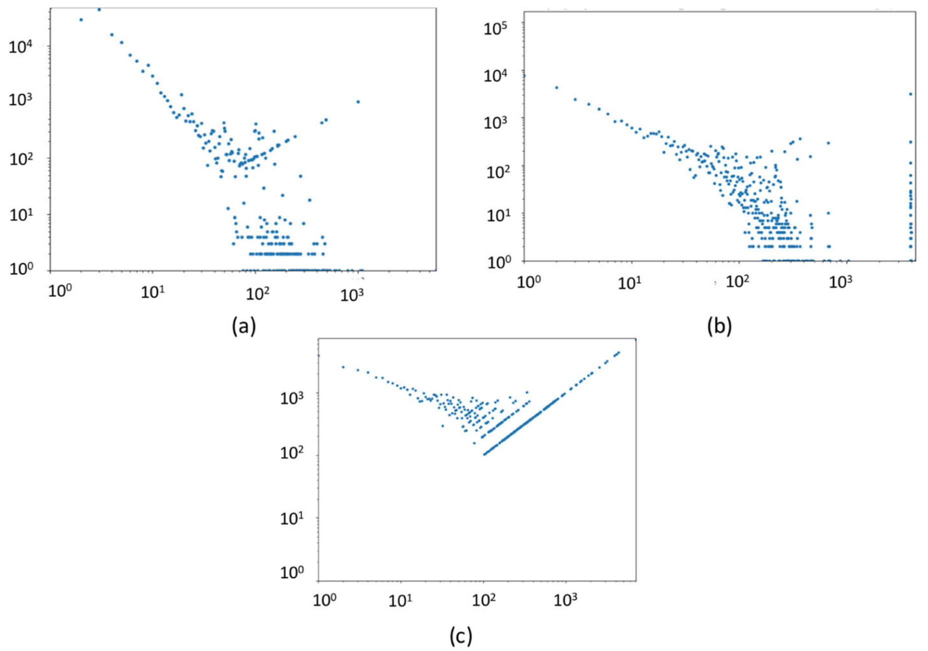

**Fig. 9** Degree distributions of higher-order networks. **a**, **b** and **c** respectively refer to networks $G_O^*$, $G_O^o$, and $G_O^i$. The x-axis is the degree, and the y-axis is the frequency of the degree

categories of officers, such as shareholders and beneficiaries, and separate networks for these different categories of officers may result in a correction of the observed deviation), which may be corrected by isolating those nodes (and their relationships) and treating them as a separate system. However, this begs the question of how to segment nodes into markets in a systematic manner. We leave this issue for future research to investigate.

Another piece of evidence that the unexpected degree distributions are likely not due to data artifacts or random noise is the power-law distributions of the connected components (Fig. 10). The plots in the figure illustrate no such unexpected behavior; they are analogous, structurally, to what we found earlier for the four networks selectively constructed earlier. The evidence leads further credence to the hypothesis that there is a fundamental difference between the phenomena expressed in the Panama Papers and in typical social and organizational networks that exhibit power law distributions.

## Summary and future work

Since their release, the Panama Papers have come under much scrutiny and study, leading to billions in recovery and collections of taxes that would not have happened otherwise, and invited a closer look at how money and corruption can manifest at a global level, involving important personages from government and industry. While not all data has been made publicly available, and such data is necessarily incomplete due to its global nature and the complex nature of entities involved, the sample that is publicly available comprises hundreds of thousands of entities and officers spread across the world, and is sufficiently representative to perform a large-scale empirical study. We perform such a study, with a focus on structure as a means of expressing and quantifying the complexity of the underlying system. We used network science as a standard scientific framework for this analysis.

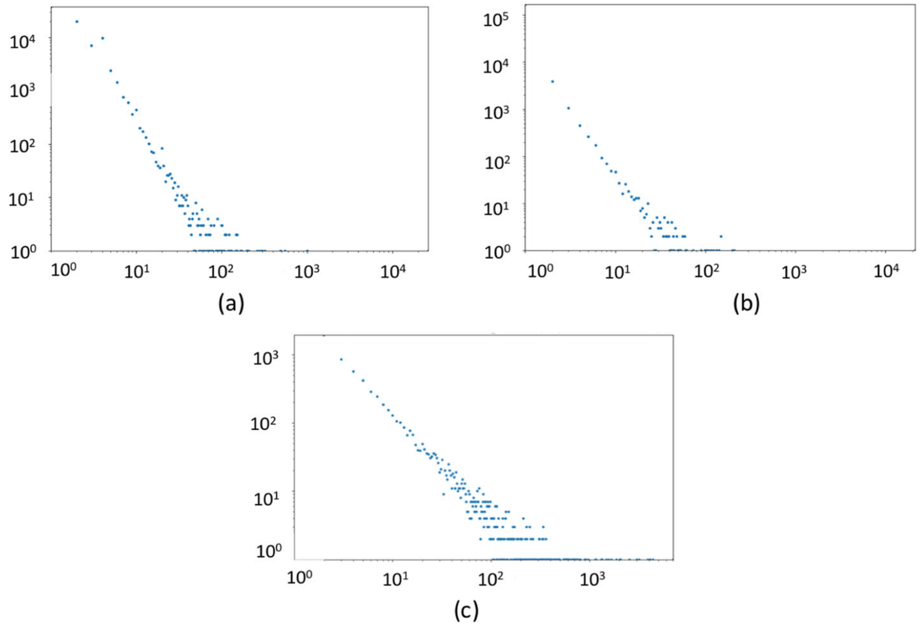

**Fig. 10** Connected component size distributions (where size is measured using number of nodes) of higher-order networks. **a**, **b** and **c** respectively refer to networks $G_O^*$, $G_O^o$, and $G_O^i$. The x-axis is the size of the component in terms of the number of nodes, and the y-axis is the frequency of the size

We derived a number of insights from these studies. Some of the findings were expected e.g., many of the first and most obvious networks that we could construct and study were found to exhibit clear power-law distributions in their in-degree and out-degree empirical frequencies, and even the exponents of some of these distributions conformed to expectations. However, other findings support the hypothesis that the Panama Papers are fundamentally different in structure than other networks of organizations, intermediaries and officers, or some combination thereof.

In particular, the lack of small-world phenomena in all selectively constructed and higher-order networks (including the officer higher-order network, which is the closest analog to an organizational social network in this domain) seems to be suggesting that a different theory is at play in this dataset, and that disconnectedness is a fundamental, emergent feature. Assuming the connected components proportionally capture some level of corrupt activity, corruption itself would seem to be a rather robust phenomenon, since targeting a few nodes would not bring the whole structure 'down', as has been found in other networks where influencers or a few highly connected and highly 'brokered' nodes play an outsize role in diffusing information and serving as a gatekeeper of sorts. Even more disturbingly, the larger components have high entropy (compared to their size) but low density, on average, suggesting that targeting or investigating nodes in the larger components would also have limited utility in cracking down on overall illicit activity. Finally, one aspect of the data that we did not cover in this article, is the dynamic nature of entities: offshore companies can quietly and quickly change ownership and beneficiaries in many jurisdictions, intermediaries can be replaced or shut down, and new shell companies can emerge, all in short order. Robustness is, therefore, an inherent feature of this system in more ways than one. We hypothesize that a theory that explains network formation and structure in the Panama Papers may shed some light on the structural and robustness properties of corruption at global levels.

We iterate that, while we certainly did not expect small-world phenomena to emerge in all networks, the fact that it did not emerge in even a *single* network, and also that the distribution of connected components (as a function of size measured in terms of number of nodes) seems to follow a clear and consistent trend throughout, suggest a similar fundamental difference compared to small-world networks. Many of the degree distributions do follow normal power-law trends, and assortativity analysis shows manifestation of slight preferential attachment, rather than any kind of rich club (or similarly assortative) effect. Ordinary preferential attachment may explain some of what we find: intermediaries that are particularly favored in certain tax havens or regions would likely draw more clients due to reputational benefits.

For some of the nodes in Fig. 9, in particular, that deviate significantly from the power-law (which are mid-degree nodes in the $G_o^*$ higher-order network, and higher-degree nodes in the $G_o^i$ higher-order network), combine quantitative or computational analyses with more qualitative analysis may be instructive for obtaining deeper insights into the operations of these networks. A sociological study of this nature is beyond the scope of this article, but is underway in our group as future work. We also hope to conduct similar analyses on the Paradise Papers and other similar datasets to assess if similar artifacts are noted in the degree distribution.

A longer-term goal is to reconcile entities in all of these different datasets, possibly after supplementing the datasets with external records and datasets, with the goal of either reinforcing or disputing the findings in this article through an analysis of the reconciled (and more complete) graph. We also plan to study the influence of intermediaries and other nodes using established centrality metrics from the literature.

### Abbreviations

USD: United States Dollars; ICIJ: International Consortium of Investigative Journalists; WCC: Weakly Connected Component; SCC: Strongly Connected Component; Max.: Maximum; tr: trace

### Acknowledgements

Not Applicable.

### Authors' contributions

Dang primarily contributed to the experiments and experimental data, while Kejriwal was responsible for writing, editing, experimental and study design, and ideation.

### Authors' information

Dr. Mayank Kejriwal (PhD, Computer Science) is a research assistant professor (Industrial and Systems Engineering), and a research lead (Information Sciences Institute) at the University of Southern California in Los Angeles, California. Dr. Kejriwal research has primarily been funded by DARPA, and involves research in AI and social network analysis for social impact problems such as human trafficking and crisis informatics. Dr. Kejriwal is a member of the Center on Knowledge Graphs at ISI, and has published papers across several AI areas. He is the author of 'Domain-specific Knowledge Graph Construction' (Springer, 2019) and an author of an upcoming MIT Press textbook on knowledge graphs as well. He has successfully transitioned work products to law enforcement for fighting human trafficking by locating suspicious activity in corpora of sex ads.
Akarsh Dang graduated with his Masters in Computer Science in May 2020 from the University of Southern California. He performed many of the experiments in this article, and conducted this research under the supervision of Dr. Kejriwal.

### Funding

This research was done pro-bono and has not received independent funding. There are no conflicts of interest.

### Availability of data and materials

The underlying data available for these studies was downloaded from https://offshoreleaks.icij.org/pages/database on Jan. 31, 2020. The version available on that date (the latest at the time of writing) is the dump from 14 Feb. 2018.

### Competing interests

The authors declare that they have no competing interests.

Received: 29 February 2020 Accepted: 30 August 2020
Published online: 11 September 2020

## References

Antal T, Krapivsky PL, Redner S (2005) Dynamics of social balance on networks. Phys Rev E 72(3):036121
Baldwin R, Forslid R, Martin P, Ottaviano G, Robert-Nicoud F (2011) Economic geography and public policy. Princeton University Press, Princeton
Barabási A-L, et al. (2016) Network science. Cambridge University Press, Cambridge
Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895
Chen P, Redner S (2010) Community structure of the physical review citation network. J Informetrics 4(3):278–290
Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703
Cooley A, Heathershaw J, Sharman J (2018) The rise of kleptocracy: laundering cash, whitewashing reputations. J Democr 29(1):39–53
Elhesha R, Kahveci T (2016) Identification of large disjoint motifs in biological networks. BMC Bioinformatics 17(1):1–18
Ellis J (2019) Corruption, social sciences and the law: exploration across the disciplines. Routledge, London
Faust K (2010) A puzzle concerning triads in social networks: Graph constraints and the triad census. Soc Networks 32(3):221–233
Forslid R, Ottaviano GI (2003) An analytically solvable core-periphery model. J Econ Geogr 3(3):229–240
Gale D, et al (1957) A theorem on flows in networks. Pac J Math 7(2):1073–1082
Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M, et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141
Graves L, Nabeelah S (2019) Gauging the Global Impacts of the "Panama Papers," Three Years Later. Reuters' Institute for the Study of Journalism, Oxford University, Oxford
Greenberg SA (2009) How citation distortions create unfounded authority: analysis of a citation network. BMJ 339:2680
Heimann F, Pieth M (2017) Confronting corruption: past concerns, present challenges, and future strategies. Oxford University Press, Oxford
Hummon NP, Dereian P (1989) Connectivity in a citation network: The development of dna theory. Soc Networks 11(1):39–63
Jancsics D (2018) Shell Companies and Government Corruption. In: Farazmand A (ed). Global Encyclopedia of Public Administration, Public Policy, and Governance. Springer, Cham. pp 4–21. https://doi.org/10.1007/978-3-319-31816-5_3566-1
Kojaku S, Xu M, Xia H, Masuda N (2019) Multiscale core-periphery structure in a global liner shipping network. Sci Rep 9(1):1–15
Krugman P (1991) Increasing returns and economic geography. J Polit Econ 99(3):483–499
Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. ACM Trans Knowl Discov Data (TKDD) 1(1):2
Li X, Chen H, Huang Z, Roco MC (2007) Patent citation network in nanotechnology (1976–2004). J Nanoparticle Res 9(3):337–352
McGregor SE, Watkins EA, Al-Ameen MN, Caine K, Roesner F (2017) When the Weakest Link is Strong: Secure Collaboration in the Case of the Panama Papers. In: 26th USENIX Security Symposium (USENIX Security 17). USENIX Association, Vancouver. pp 505–522. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/mcgregor
Miethe J, Menkhoff L (2017) Dirty money coming home: Capital flows into and out of tax havens, Annual Conference 2017 (Vienna): Alternative Structures for Money and Banking 168082, Verein für Socialpolitik / German Economic Association
Miller S (2017) Institutional corruption: a study in applied philosophy. Cambridge University Press, Cambridge
Milo R, Kashtan N, Itzkovitz S, Newman ME, Alon U (2003) On the uniform generation of random graphs with prescribed degree sequences. arXiv e-prints. https://ui.adsabs.harvard.edu/abs/2003cond.mat.12028M
Moody J (1998) Matrix methods for calculating the triad census. Soc Networks 20(4):291–299
Nerudova D, Solilova V, Litzman M, Janskỳ P (2020) International tax planning within the structure of corporate entities owned by the shareholder-individuals through panama papers destinations. Dev Policy Rev 38(1):124–139
Neu D, Saxton G, Everett J, Shiraz AR (2020) Speaking truth to power: Twitter reactions to the panama papers. J Bus Ethics 162(2):473–485
Newman ME (2003) Mixing patterns in networks. Phys Rev E 67(2):026126
Obermayer B, Obermaier F (2016) The panama papers: breaking the story of how the rich and powerful hide their money. Oneworld Publications, London
O'Donovan J, Wagner HF, Zeume S (2019) The value of offshore secrets: evidence from the panama papers. Rev Financ Stud 32(11):4117–4155
Rabab'Ah A, Al-Ayyoub M, Shehab MA, Jararweh Y, Jansen BJ (2016) Using the panama papers to explore the financial networks of the middle east. In: 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, Barcelona. pp 92–97
Radon J, Achuthan M (2017) Beneficial ownership disclosure: the cure for the panama papers ills. J Int Aff 70(2):85–108
Rotberg RI, Carment D (2018) Canada's corruption at home and abroad. Routledge, London
Thomas DR, Pastrana S, Hutchings A, Clayton R, Beresford AR (2017) Ethical issues in research using datasets of illicit origin. In: Proceedings of the 2017 Internet Measurement Conference. ACM, London. pp 445–462
Trautman LJ (2016) Following the money: lessons from the panama papers: part 1: tip of the iceberg. Penn St L Rev 121:807
Tuttle H (2016) Data deluge: what the panama papers leak means for business. Risk Manag 63(6):22
Wasserman SS (1977) Random directed graph distributions and the triad census in social networks. J Math Sociol 5(1):61–86
Watts DJ (2004) Six degrees: the science of a connected age. WW Norton & Company, New York

Wiedemann G, Yimam SM, Biemann C (2018) A multilingual information extraction pipeline for investigative journalism. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium. pp 78–83

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.