Coscia et al. *EPJ Data Science* (2018) 7:12
https://doi.org/10.1140/epjds/s13688-018-0141-0

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Mapping the international health aid community using web data

Michele Coscia[1,2*], Katsumasa Hamaguchi[3], Maria Elena Pinglo[3] and Antonio Giuffrida[3]

*Correspondence:
michele_coscia@hks.harvard.edu
[1]Naxys Department of Mathematics, University of Namur, Namur, Belgium
[2]Center for International Development, Harvard University, Cambridge, USA
Full list of author information is available at the end of the article

Abstract

International aid is a complex system: it involves different issues, countries, and donors. In this paper, we use web crawling to collect information about the activities of international aid organizations on different health-related topics and network analysis to depict this complex system of relationships among organizations. By systematically collecting co-occurrences of issues, countries, and organization names from more than a hundred websites, we are able to construct multilayer networks describing, for instance, which issues are related to each other according to which organizations. Our results show that there is a surprising amount of homophily among organizations: organizations of the same type (multilateral, bilateral, private donors, etc.) tend to be co-cited in groups. We also create a taxonomy of issues that are generally mentioned together. Finally, we perform simulations, showing that messages originating from different organizations in the international aid community can have a different reach.

Keywords: Web crawling; Global health architecture; Complex networks

1 Introduction

International aid—or "Global Health Architecture" [1] when focusing on health topics in particular—is a complex system [2]. Aid involves the voluntary transfer of resources from one country to another, with the aim of addressing an emergency or a systemic problem in the recipient country. There are many issues that need assistance, each affecting at different levels many possible recipients, and many donors, donor's interests, and donor types—governments, multilateral organizations, charities, private citizens. To find the most effective way to invest in aid is difficult. To make matters worse, this complexity renders difficult for organizations to perform strictly under their own mission statement, and for donors to track whether their investment is used as intended. In other words, there is a disconnect between *de jure* and *de facto*—between theory and practice.

In this paper, we address the problem of making sense of the global health architecture. Specifically, we focus on the subset of international aid centered on health. We do so by using the tools of complexity science. Since international aid is a complex system, it can be modeled and studied as such. Taking this perspective enables us to understand its emerging properties. We want to know who are the leaders of the international communities in the online world, and which are the issues and the places on which they focus most of their attention—again limiting ourselves to Web public speech.

To take such perspective, the most crucial problem to solve is finding an appropriate data source. Following previous work [3], we decide to use the speech each aid organization uses in their public communications, specifically their websites. Websites have several advantages over other forms of speech. First, they are easy to collect: differently from conferences or plain text documents, websites are made to be machine readable. Second, they provide rich and structured data: the Web has been developed to allow explicit connections between documents, that can be used to translate the data directly into a complex network, which is a useful tool to understand complexity [4–6].

Using web crawling to reconstruct the emerging relationships between entities is not a new idea. This approach has been tested in numerous scenarios, from reconstructing social networks [7] to associations between entities from news items [8, 9], from place relatedness [10] to linking politicians and policy finance [11], and relationships between health-research institutions [12].

Following [3], we collect the content of the websites of more than a hundred international aid organizations, working on issues related to health. We collect the pages containing at least one among a specific set of keywords and countries. We consider that the relative number of times an organization mentions an issue or a country is directly proportional to its focus on that issue/country. At the same time, the number of occurrences of the organization names with an issue/country in other organizations' websites is a signal of how much the organization is seen as related to that issue/country. This effectively uses the discourse of peers—the international aid organizations—to build a map of associations between issues, countries, and organizations.

We differ—and improve over—[1, 3] in several ways. First, we perform an actual web crawl. While in previous work we relied on the content indexed by Google, here we collect the content first hand. This solves a number of issues related to unreliable occurrences counts from Google [13, 14]. Second, we move from using tripartite networks to modeling our data with multilayer networks. Multilayer networks are networks in which the same set of nodes can connect in different layers, which represent different types of connections [15–17]. For instance, we create an issue multilayer network by connecting together issues that are frequently co-mentioned in the same web pages. Each layer contains connections coming only from a single website. Therefore, we can compare the topology of the issue–issue network as seen by different organizations.

We can summarize the paper's contributions as follows. First, we delineate an improved approach for the construction of international aid complex networks from web crawled data. Second, we are able to unveil relationships between organizations, countries and issues. In organizations, we discover a strong homophily. Homophily in social systems is the tendency of connecting to people who are similar to us [18]. In our network, this means that multilateral organizations tend to be co-cited together, and this holds for all other types of organization. Countries cluster geographically. The co-citation network of issues is useful to create an issue taxonomy, highlighting which issues are considered related by the aid community. Third, we analyze relationships between unlike entities, for instance which issues are the most (or least) related to a given organization.

We strengthen our analysis with a case study focused on the World Bank. We choose the World Bank because we have additional information about its inner mechanics. In the case study, we show that the organization-issue and organization-country relatedness measure from web data correlates significantly with the actual amount of effort the World Bank

devotes to issues/countries. We also study the reach that a message originating from the World Bank can have in the organization citation network. By using a simple Susceptible-Infected model [19], we can simulate the spread of information through the international aid network.

The collected data and the code necessary to replicate our experiments are freely available.[a]

2 Methods

The methods of this paper are divided in four phases. The aim is to go from a set of topics, countries, and organizations to the creation of a complex multiplex network, informing how these entities relate to each other. The four phases are:

1. *Entity Definition*, which aims at uniquely identifying the data sources and nodes of the networks.
2. *Web Crawling*, whose task is to gather the raw information from the webpages of the data sources.
3. *Building the Networks*, in which we show how to clean the noise from the raw crawled data.
4. *Postprocess Blacklisting*, where we use the clean data to identify which entities we should drop due to ambiguities that could not be ruled out in the first phase.

2.1 Entity definition

The first step for extracting relations between entities from web pages is to define which are our entities. Entities are coded through keywords. Here we focus on three types of entities: the aid organizations part of our study, the issues related to health and aid, and the set of countries in which the organizations operate. All lists of entities have been curated by domain experts.

We start from the latter mentioned: the countries. This is arguably the less problematic task among the three, because countries are generally identified by their name. Thus, the list of country entities is simply the list of country names on which we want to focus. There are few exceptions, for instance in the case of Jordan we make sure the webpage is writing about the kingdom of Jordan and not about a person called Jordan. In the case of the Kyrgyz Republic, we have to be aware that, in some cases, it is referred to as "Kyrgyzstan". Table A1 reports the list of countries that were selected in this study as examples, to illustrate how the organizations relate in specific country-settings based on variety in income level, fragility status, and size of the World Bank Group support to health services. The countries were selected among the subsample of countries that had received World Bank and IFC financial support to improve health services during the 2005–2016 period, in order to cover all World Bank Group geographical regions; different income levels and fragility.

Next, we define the international aid organizations that are part of the study. This list has two ends: the aid organizations are the subject of the study, but at the same time they are the source of the data of the study. It is necessary not only to refer to them in an unambiguous way, but also to identify their website. In many cases, it was possible to name an agency which did not have a website, or it did not have a website in English (more on this requirement in the next section), and thus the organization was dropped from the study.

For each organization, we scan the webpages in search of its full name. Since it is rare that organizations are mentioned by their full name—in some cases it can be quite long—we

look also for its official acronym. Moreover, we define, where appropriate, the colloquial and most common handle for the organization: for instance the "Bill & Melinda Gates Foundation" is usually mentioned simply as the "Gates Foundation".

There are some cases in which an organization can have multiple websites. For instance, the World Health Organization has five of them. In that case, all websites are crawled and the information gathered from all of them is connected to WHO. Table A2 reports the list of countries on which we focus in this study, along with their website's URL.

Additionally, each organization can belong to an organization class. There are seven classes:

- Bilateral (DAC): a bilateral is a public organization established by a single country which interfaces directly (bilaterally, hence the name) with its recipients. In this class we consider only those organization from countries that are part of the OECD's Development Assistance Committee (DAC). Examples: UK's DfID, Japan's JICA, USAID.
- Bilateral (non-DAC): same as before, only considering countries which are not part of OECD's DAC. Examples: Russia's Rossotrudnichestvo, Kuwait Fund, Malaysia's MTCP.
- UN System: autonomous organizations working with the United Nations and each other through the coordinating machinery of the United Nations Economic and Social Council at the intergovernmental level, and through the Chief Executives Board for coordination (CEB) at the inter-secretariat level. Examples: FAO, ILO, UNICEF.
- Multilateral: Multilateral organizations obtain their funding from multiple governments and spend it on projects in various countries. Examples: World Bank, IMF, African Development Bank.
- NGO: acronym that stand for "Non-Governmental Organization", these are leading international non-profit, voluntary citizens' groups providing supporting health services in less developed countries. Examples: Save the Children, Doctors without Borders, Red Cross.
- Philantropy: similar to NGOs, these are leading non-profit non-governmental entities that utilize donated assets and income to support health services in less developed countries. Examples: Gates Foundation, MasterCard Foundation, Rockefeller Foundation.
- Health Initiative: programmatic global partnerships or initiatives in which the participants: (i) dedicate resources toward achieving agreed objectives over time; (ii) conduct activities that are global, regional, or multicountry in scope; and (iii) establish a new organization with shared governance and management unit to deliver these activities. Examples: Power of Nutrition, the Vaccine Alliance, Stop TB.

Finally, we define the issues we want to study. In this paper, we are focusing on health issues and organizations, so our keywords are narrowed to this domain. Like in the case of countries, we selected keywords that characterize World Bank Support to health services provided during the 2005–2016 period. Each keyword is reduced to lower case. This means that the crawler will match a page with a keyword regardless of its case.

While also the organization names are case insensitive, this does not apply to the acronyms, which have to match also the case. This is to avoid to match "WHO" to the pronoun "who". This measure is necessary because we do not perform standard stopwording to the text, meaning that we do not remove syntactic words (like "the", "and", etc.) from the

webpage text. The reason is that organization names contain stopwords, and we would miss matches if we removed them.

A domain expert defined the keywords to avoid ambiguity. For instance, the simple word "health" is not a keyword, because the word can be used ambiguously as a metaphor (e.g. in the formula "the health of a country's economy"). Table A3 reports the list of keywords on which we focus in this study.

2.2 Web crawling

The data is collected by a standard web crawler, implemented in Python using the scrapy package.[b] Table A2 lists the starting URLs, which are the input of the crawler. The crawler is instructed never to leave the starting domain, so no page outside the selected websites is considered. If a domain has a sub-domain (e.g. www.worldbank.org contains data.worldbank.org) the sub-domain is also crawled. We impose no depth limit, meaning that the domains are fully crawled, without discarding pages that are more than n clicks away from the home page.

In some cases, webpages need JavaScript to properly render their content and hyperlinks to other pages. We use a headless web browser—PhantomJS[c]—able to execute the code to render the page.

Each page is firstly tested against the list of our issues, as defined in Table A3. If the page does not contain any of the specified issues, it is discarded and not used for any of the analysis in this paper. If the page does contain at least one issue, then we store all the keywords—issues, countries and organizations—mentioned, and all hyperlinks to other websites included in the study.

There are a number of challenges related to peculiarities of data coming from websites. We focus on the following three:

1. Issues, organizations and countries might appear in menu items. This means that a co-appearance count might be inflated by pages having a country name in the navigation menus but not in the actual content.
2. The crawler uses the URL of a page to identify it as unique. In some cases, the same page might have different URLs, due to parameters appended to the URL which do not change the content of the page.
3. Some organizations have minimalistic websites containing a general overview of their activities, while other organizations might have multiple pages for every specific project/grant.

In all three cases, we would obtain inflated counts for term co-appearances. Our counter-measures are as follows.

Regarding issue #1, an easy solution would be to automatically filter navigation menus in headers/footers. However, given differences and lack of standards in web design, this is not feasible when crawling multiple websites. We are then forced to perform a post-process of the crawled data. We observe that, if terms are appearing in structural components of a website, they will always appear the same number of times in each page. In other words, menu items will result in a background occurrence frequency for some keywords. To take this into account, we establish the mode appearance of a word in a website. If a page mentions the word less than or equal to the website mode, the mention is discarded, unless the mode happens rarely, thus it cannot be associated with a menu item.

Regarding issue #2, we also propose a post-process cleaning operation. If the same page was crawled n times from n different URLs, then we will find n copies of the same sets

of terms, in the same order and with the same frequency. We scan all the crawled pages from a website and, if this is the case, the n copies of the pages will be counted only as one. This is an imperfect solution, because there could be two genuinely different pages with the very same term composition, but we assume that such coincidence will be rare and unable to affect significantly our results.

We address issue #3 in the next section, because its solution is implemented in the process that transforms the crawled webpages into the edges of the networks we build.

2.3 Building the networks

Once the web crawling is complete, the result is a set of pages—cleaned with the post processing procedure explained in the previous section. A page p is defined as a set of entities that co-appear on the page, among which at least one has to be a health issue, otherwise the page is discarded. Every time two entities—say i and j—co-appear in a page of a website w, we increase their co-occurrence counter n_{ij}^w by one. We keep a different co-occurrence counter for each website to address the issue #3 mentioned in the previous section.

We transform the co-occurrence counts from every website such that each additional co-occurrence provides diminishing returns to the association strength between the two terms. In other words, if terms i and j co-appear in website w n_{ij}^w times, then we define their connection strength from w as $\log_{10}(n_{ij}^w)$. When we aggregate across all websites $w \in W$, we define their overall "Web Score" WS_{ij} as:

$$WS_{ij} = \sum_{w \in W} \log_{10}\left(n_{ij}^w + 1\right).$$

WS_{ij} still provides a higher connection strength from w if w has many pages mentioning i and j, but it is difficult for it to dwarf smaller websites that could mention i more often with a third entity k rather than with j. Note that WS_{ij} is symmetric, meaning that $WS_{ij} = WS_{ji}$. This is because co-appearance is a symmetric relationship ($n_{ij}^w = n_{ji}^w$).

In this paper we also look at a citation network, connecting organization i to organization j if i cites or links j. The connection strength is still estimated using WS_{ij}, replacing n_{ij}^w with c_j^i. Rather than recording the number of co-appearances of i and j, c_j^i records the number of times i mentions or links j in its website. In this case, symmetry is broken, as $c_j^i \neq c_i^j$: even if i links to j, it is possible that j does not link back.

WS_{ij} is at the basis of all our networks with a single layer (i.e. the ones we aggregate from the entire crawled data): if i and j are entities of the same type (say they are both countries), then the network is unipartite. If i and j's type is different—say one is an organization and the other a issue—then the resulting network is bipartite.

We can create a version of WS_{ij} for generating multilayer networks. If we focus on a single country c by selecting exclusively the pages mentioning c, we can create a derived WS_{ij}^c. If we generalize this by collecting all possible WS_{ij}^c $\forall c \in C$, the result is a multilayer network whose layers are the network for a given c, which represent how that particular country is "seen" by the international aid community. The same operation can generate multilayer networks whose layers are co-occurrences coming from a single website w, which represent how that particular organization "sees" the international aid community on health.

The last step in the network generation process is network backboning. Web data is very noisy, meaning that a page might mention two entities in very different parts of it, and the

relation between the two might be weak. Network backboning is a cleaning procedure designed specifically to identify and remove from the network spurious communities. For most of our networks, we use the noise-corrected backbone, which establishes the expectation for an edge weight and keeps it in the network only if the actual edge weight is higher than this expectation [20]. This method is specifically designed for count networks, i.e. networks whose edge weights are results of a discrete counting process. This is the case for the web networks studied here, as each co-appearance is increasing a counter.

The citation network is a special case for two reasons. First, citing and linking a website is an act less prone to noise: a hyperlink is a stronger signal than the co-appearance of two terms in a text. Second, it is a network whose higher-order is more meaningful than in the co-appearance network. The number of edges required to go from i to j in a citation network is meaningful, because it is proportional to how difficult it is for information to flow from i to j. If two entities are not connected in a co-appearance network it means they are never mentioned together, so it makes little difference if their network distance is two or four edges. For these reasons, this backbone is instead generated using the high-salience skeleton [21], which is built exactly looking at shortest paths in the network.

2.4 Postprocess blacklisting

Even if we use proper names and acronyms to distinguish an organization's name from the rest of a page text, our procedure is not exempt from possible mistakes. Names like "World Bank" or "Plan International" are phrases that can be used colloquially, without necessarily referring to those organizations in particular. We create a procedure to blacklist organizations in case our pre- and post-process were unable to perform this distinction properly.

For this phase, we use specifically the citation information. Each organization can be cited in two ways: (i) in the text of a page and (ii) by a hyperlink. While the former way is affected by the aforementioned problem, the latter is not: a link cannot be used colloquially in a text without wanting to explicitly refer to that particular website. Therefore, if an organization is disproportionally cited through mechanism (i), then we can argue that those documents were using the sentence without really referring to the organization itself. Those organizations are likely to introduce noise in the networks and have to be dropped.

Our test is performed using a simple linear regression, given the assumption of the strong relation between mechanisms (i) and (ii). In practice, we want to use the number of times an organization is cited by hyperlink to predict how many times it will be cited by name. Organizations with a high residual are candidates to be dropped from our networks—although they are kept as data sources, since the content of their website is not affected by this issue.

When we look at the top residuals of this regression, there is one organization that stands out: Plan International. Its residual is 68% higher than the second largest residual. The average of this ratio for the other residuals in the top 10 is only 12%. As a consequence, the only organization we drop from our networks after this procedure is Plan International.

3 Results

3.1 Unipartite networks

We start our analysis by looking at the unipartite networks. The first on which we focus is the network of co-citations of international aid organizations. In this network, two orga-

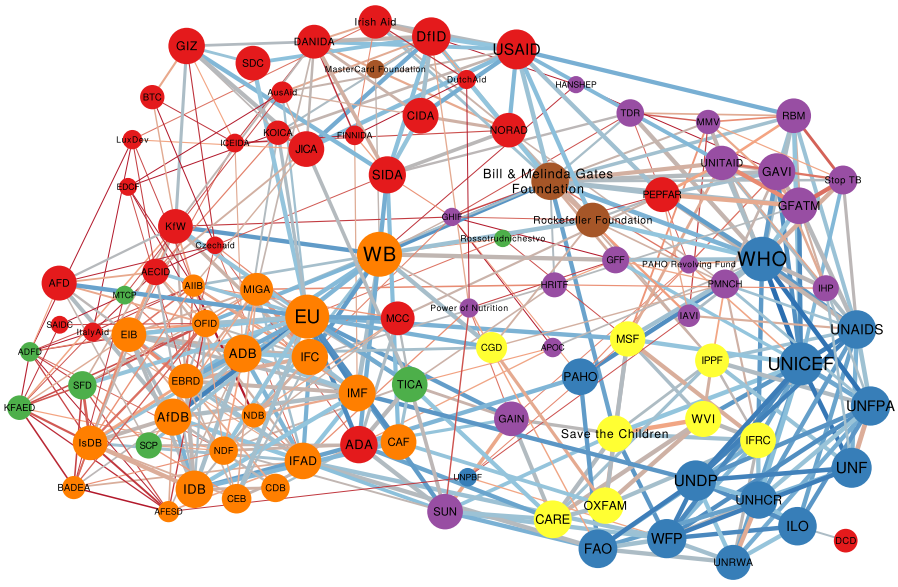


Figure 1 The organization co-citation network. Each node is an international aid organization. The node's size is relative to its aggregate *WS* value. The node color reflects the organization type: red = DAC (Development Assistance Committee) bilateral; blue = United Nations; green = non-DAC bilateral; purple = health initiative; orange = multilateral; yellow = Non-Governmental Organization; brown = private philanthropy. Edges are sized according to their *WS*, and colored according to their significance

nizations are linked to each other if they are mentioned in the same web page a significant amount of times.

Figure 1 depicts this network. The first noticeable property of the network is the tendency of organizations to cluster with similar organizations. This homophily can be estimated by calculating the modularity of the partition obtained by classifying organizations according to their class (as described in Sect. 2.1). Modularity is a measure estimating how many more edges are inside a node community than we would expect [22]. Modularity takes values between –1 and 1. In this case, we observe a value of 0.34. This is remarkably high considering that we obtain a modularity of 0.47 when we detect node communities with an actual community discover algorithm (Infomap [23]). This means that the organization class is a reasonable way to partition this network, i.e. organizations of similar classes tend to cluster together.

Note that this is a significance map, meaning that we only see the associations that are stronger than expected. There are strong associations that are not in this network, because they are a reflection of business-as-usual for the connected organizations. The distribution of edge weights is broad, and we depict it in Fig. 2. The distribution has a strong exponential cutoff. This means that very well cited organizations fail to be mentioned together as much as we would expect—as this would generate a power-law. Table 1 reports the top ten organizations associated with the World Bank according to the *WS* measure.

Keeping our attention on the organization-organization network, we now focus on its citation variant: here organizations are connected not if they co-appear in a page, but if they directly mention (or link) one another. In this case, connections are not symmetric: the edges have a direction going from the citer to the citee.

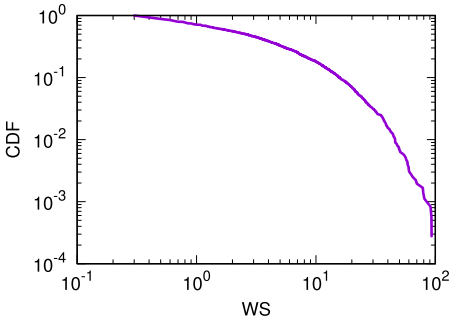


Figure 2 Cumulative distribution of *WS* per pair of organizations in the co-citation network

Table 1 Top ten organizations according to *WS* for the World Bank in the co-citation network

Rank	Organization	*WS*
1	WHO	92.69
2	European Union	91.07
3	UNICEF	78.52
4	UNDP	69.99
5	USAID	58.23
6	IMF	56.66
7	UNF	53.71
8	AfDB	49.82
9	DfID	49.57
10	Gates Foundation	48.74

Table 2 The ten most central organizations in the citation network, and their incoming citation score

Rank	Organization	Betweenness	In-*WS*
1	WB	0.193	149.18
2	UNDP	0.115	104.01
3	EU	0.091	147.25
4	FAO	0.065	81.54
5	UNICEF	0.053	126.44
6	ADB	0.039	67.73
7	WFP	0.038	86.91
8	WHO	0.033	146.15
9	IMF	0.027	68.56
10	AfDB	0.026	65.73

The first thing we are interested in investigating is which organization is the most cited and which one is the most central in this kind of structure. Table 2 reports the ten most central organizations in the network, according to betweenness centrality [24]. Betweenness centrality tells us the fraction of paths in the network that would get longer—or disconnected—if the node were to be removed from the network.

Note that being the most central node in the network does not necessarily mean that the node is the most cited. In fact, the third most cited organization (according to the sum of the *WS* of the edges pointing to it) is the World Health Organization, which is eighth most central. UNDP and FAO received less citations, but they have higher betweenness centrality. This means that the additional citations come from either the same edge or from other organizations that do not contribute to the centrality of the WHO.

In the citation network, another important question regards connected components: is there a citation path going from an organization to any other organization? A disconnected

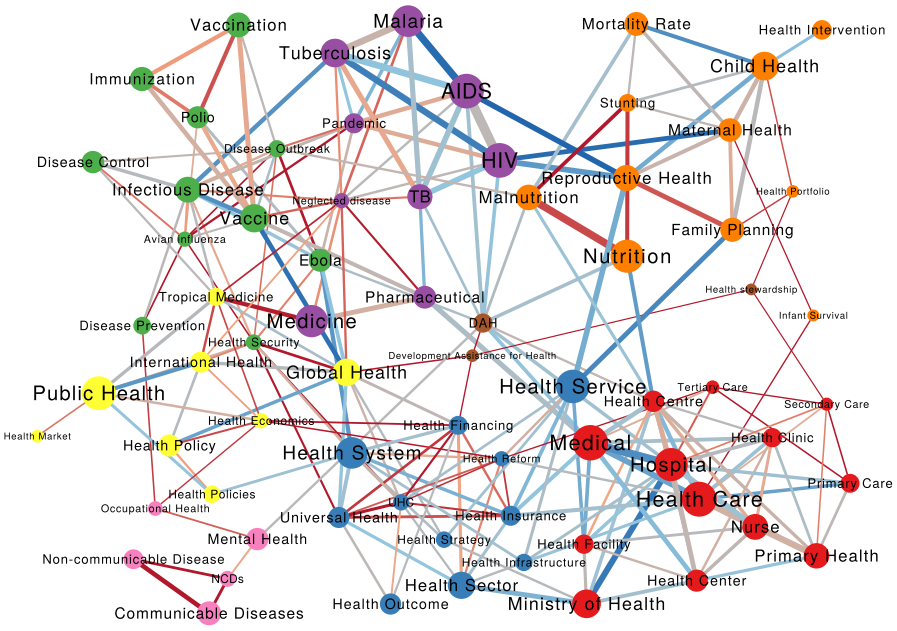


Figure 3 The issue co-citation network. Each node is a health issue. The node's size is relative to its aggregate *WS* value. The node color reflects the network community. Edges are sized according to their *WS*, and colored according to their significance

space would mean that there is information that cannot reach every organization in the system.

The citation network has only one weakly connected component. This means that there is a path between any two organizations, provided that we ignore the direction of the edge [25]. The majority of the network is composed by nodes that are target of a citation, but do not cite back in sufficient amount to clear the backboning threshold. There is a core of 15 aid organizations part of the only non-trivial strongly connected component (i.e. a component containing more than one node). With the exception of Save the Children, all organizations in Table 2 are part of this core of the network.

We conclude the section on unipartite networks by looking at the other two types of networks we generated: the issue–issue and country-country networks. We do not perform analysis on them as this paper is more focused on aid organizations. However, they are objects with interesting descriptive power. They tell us how aid organizations talk about health issues and countries.

Figure 3 depicts the issue network. Two issues are connected if they co-appear in webpages more often than we would expect given their popularity. There are two interesting aspects. First, terms that are synonyms or quasi-synonyms are connected in this network—such as "DAH" and "Development Assistance for Health", or "TB" and "Tuberculosis". This is a sign that the methodology is valid, as it was able to catch such relationships.

The second aspect concerns our ability of defining clusters of issues. We perform a simple community discovery with the Infomap algorithm [23] and it returned us the clusters depicted in the figure. The communities help us in making sense of different issues as being part of macro categories. For instance, we can delineate a reproductive health macro

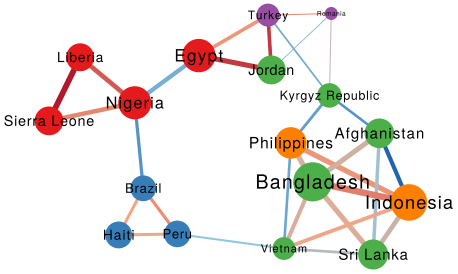


Figure 4 The country co-citation network. Each node is a country. The node's size is relative to its aggregate *WS* value. The node color reflects the world region in which the country is located. Edges are sized according to their *WS*, and colored according to their significance

category (in orange) as distinct from the one focused on infrastructure and governmental cooperation (in red).

Similar considerations can be made for the country network. Figure 4 depicts it. This network can also be used as a mean of validation of our Web data. We expect to find connections between countries that share a similar geographical and cultural position. Countries affected by the same issues should also be closely connected.

In fact, we see that there is a strong geographical component, showing an African, American, Middle East and Far East cluster. Romania is the only European country, and thus has smaller connections. Yet, its most significant connection is with its neighbor Turkey.

3.2 Bipartite networks

We now turn our attention to bipartite networks. In a bipartite network, we have two classes of nodes. Edges only connect nodes belonging to unlike classes. In this work we have three possible types of bipartite networks: organizations connected to keywords, organizations connected to countries, and countries connected to keywords. The connection criterion is the same used for unipartite networks: the co-appearance of the two terms in the same webpage.

The first network we analyze is the one connecting organizations to keywords. Our first question is: how can we estimate which issues are the most associated with an organization? To answer this question we create a derived *WS* measure. WS_{oi} is the total number of co-appearances of organization o and issue i. We can define WS_i as the total number of times i was mentioned. Note that $WS_i \neq \sum_{\forall o \in O} WS_{oi}$, because we can have a page mentioning i and two different organizations, so that page would be double counted.[d]

The ratio between WS_{oi} and WS_i is informative: it is the share of pages mentioning i that also mentioned o. The higher this share, the more o is seen as relevant to i by the aid community—even if o might not mention i at all. Table 3 reports the top and bottom issues according to this ratio for the World Bank. Note that the World Bank is mentioned more often together with AIDS and HIV than it is mentioned together with Infant Survival. However, HIV/AIDS are topics discussed very broadly and very often, and only in around 4% of the times the World Bank is mentioned. This is still higher than expectation of uniform mentions: the expectation is lower than 1%, given that we have more than 100 organizations in our sample.

We can visualize the relationship between the World Bank and all issues on the issue network we built in the previous section. In Fig. 5 (left), we use the same layout used in Fig. 3. The only difference is in the color of the nodes. Instead of showing network communities, we show the value of $WS\%$ ($= 100 * WS_{oi}/WS_i$). From the figure, we can see

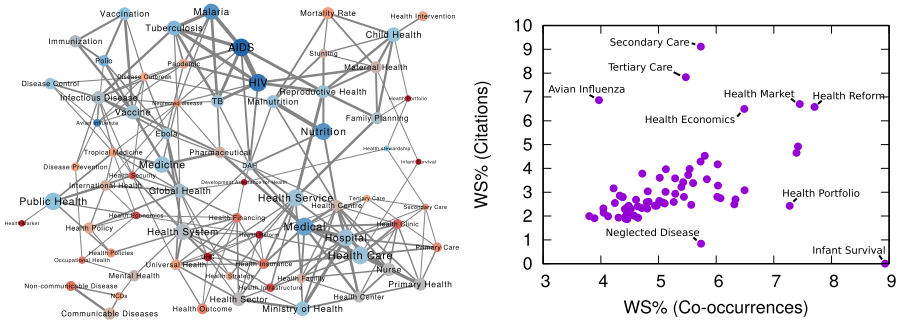


Figure 5 (Left) The intensity of World Bank co-appearances according to *WS*%, from high (red) to low (blue). (Right) The relationship between co-appearances (*x*-axis) and direct mentions (*y*-axis) of World Bank for all issues

Table 3 The ten most associated issues to the World Bank according to *WS*

Rank	Keyword	*WS*%
1	Infant Survival	8.93
2	Health Reform	7.71
3	Health Market	7.45
4	UHC	7.42
5	Development Assistance for Health	7.39
…	…	…
66	Nutrition	4.05
67	Malaria	4.04
68	Avian influenza	3.97
69	HIV	3.89
70	AIDS	3.80

that the World Bank tends to have lower *WS*% only for very popular issues (large nodes), and it tends to be mentioned disproportionally often with smaller issues.

Is there an alignment between what the World Bank is talked about (i.e. the terms with which it co-appears) and what the World Bank is talking about (i.e. the terms it mentions in its website)? Fig. 5 (right) is an attempt to answer this question. The *WS*% indicator as calculated so far is equal to WS_{oi}/WS_i. If we substitute the numerator, we can estimate the share of pages mentioning i coming from a given organization website o: WS^o_i/WS_i. In a perfectly aligned world, these two ratios should be the same: an organization is mentioned together with an issue exactly as much as this organization itself mentions the issue. Disagreements in these two ratios imply that o thinks some other issue is more relevant to itself than i.

Figure 5 (right) shows that there is a correlation between the two ratios, but it is not perfect. The vast majority of issues is clustered into a clear linear relationship; however, there are a handful of outliers. Secondary/Tertiary Care and Avian Influenza are mentioned by the World Bank more than the World Bank is mentioned with them; on the other hand, the World Bank mentions Infant Survival and Neglected Disease proportionally less than it is mentioned about them.

A plausible reason to explain why in the World Bank site "Infant Survival" is found more often with other partners (i.e. high co-appearances relative to citation), while other terms, such as "Avian Flu" is found more often in isolation (i.e. high citation relative to co-appearances) could be that the World Bank operates to improve "Infant Survival" often

Table 4 The ten most associated countries to the World Bank according to *WS*

Rank	Keyword	*WS*%
1	Vietnam	4.33%
2	Kyrgyzstan	4.33%
3	Haiti	4.27%
4	Turkey	4.24%
5	Peru	4.24%
…	…	…
13	Sierra Leone	4.14%
14	Sri Lanka	4.11%
15	Indonesia	4.11%
16	Philippines	4.08%
17	Nigeria	4.02%

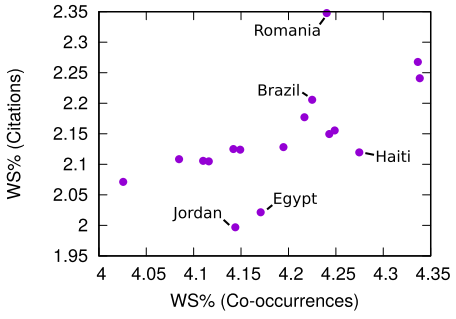


Figure 6 The relationship between co-appearances (*x*-axis) and direct mentions (*y*-axis) of World Bank for all countries

with other health organizations (e.g. UNICEF and WHO). On the other hand, the World Bank often operates in the "Avian Flu" domain with no-(human)health organizations (e.g. OIE—World Organisation for Animal Health) as avian flu is relevant both to human and animal health.

We can repeat the same exercise in the organization-country bipartite network. Table 4 reports the results of the same analysis performed for Table 3; Fig. 6 depicts the same scattergram shown in Fig. 5 (right), shifting the attention from issues to countries.

The main difference between the two bipartite networks is that *WS*% distributes differently. There is less variation in the organization-country network than in the organization-keyword network. For the World Bank, the difference between the most associated country (Vietnam) and the least associated country (Nigeria) is just above 0.3 percentage points.

Just as in the case with issues, also for countries there is a strong correlation between the number of times the World Bank is mentioned about a country and the number of times it mentions the country itself—with the presence of some outliers. Romania is mentioned by the World Bank more than we would expect given the speech of the rest of the aid community, while Jordan and Egypt are mentioned fewer times.

We conclude this section by looking at the country-issue bipartite network. In this case we cannot compare the association as emerging from co-occurrences against direct mentions, because we do not have a source of data on what a country talks about. For this reason, we limit ourselves to providing a simple ranking. We want to answer the question: which issue is most related to which country?

Table 5 provides the answer. There is only one issue that ranks as the top associated concern in more than two countries and that is Infant Survival. From Fig. 3 we could see that

Table 5 The most associated issue per country to *WS*

Issue	Country
DAH	Brazil
	Sri Lanka
Ebola	Liberia
Health Market	Peru
Health Portfolio	Afghanistan
Health Reform	Jordan
Health stewardship	Vietnam
NCDs	Kyrgyzstan
Nurse	Haiti

Issue	Country
Infant Survival	Sierra Leone
	Nigeria
	Indonesia
	Philippines
	Bangladesh
	Egypt
Secondary Care	Turkey
Vaccine	Romania

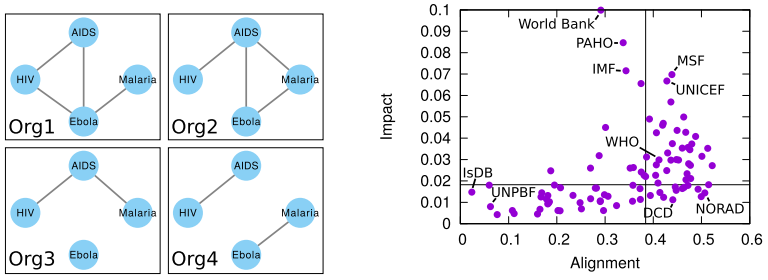


Figure 7 (Left) A toy example showing the fundamental structure of the multilayer network we analyze. ach quadrant is a different layer, containing the same nodes connected by different edges according to the data coming from the websites of different organizations. (Right) The relationship between the Alignment (*x*-axis) and the Impact (*y*-axis) of each organization in the multilayer network. The black lines show the median values of the two measures

this issue is not mentioned often in the crawled websites. However, most of its mentions come from a collection of countries: the six countries listed in the table make up for more than 77% of all mentions of the issue.

3.3 Multilayer networks

In this section we focus on multilayer networks, i.e. networks whose nodes can be connected with different criteria. We focus on a single variant on all the possible multilayer networks we could build. Analysis similar to the one delineated in this section can easily be carried on the other types of multilayer networks. Here we are interested in investigating the network connecting issues together.

The layers of the network come from the different websites crawled. In other words, we build a unipartite issue–issue network just like we did in the Unipartite Networks section, by looking exclusively at data coming from a website at a time. In this way, we have a layer showing us the issue–issue network in the eyes of the World Bank, a layer for UNICEF, a layer for the WHO, and so on. Figure 7 (left) shows an example of such structure.

In the Unipartite Networks section we saw that there is an issue–issue network arising from the combination of the discourse of all international aid organizations. From that network, we discovered that—for instance—HIV is related to Maternal Health. From where is this connection coming? How do the networks emerging from the speech of each aid organization combine to form the big picture shown in Fig. 3?

In [17], the authors introduced two concepts that are useful for thinking about this question. These concepts were “complementarity” and “redundancy”. If we look at all the edges

in the multilayer network, we can record in how many layers a connection is reproduced. If on average all connections are present in most layers, then we can conclude that there is a high degree of redundancy: the deletion of an edge in a specific layer does not affect the connectivity of the multilayer network as a whole. However, if most edges are present in a single layer, then the deletion of a single edge from i to j in the only layer in which they are connected has a dramatic effect. If we perform such deletion, there is no direct way to go from i to j.

We can redefine redundancy ρ here to apply it to a single edge. The redundancy ρ_{ij} of the edge connecting i to j is simply the number of layers in which the edge appears, divided by the total number of layers in the network. If l is a layer containing a set of edges from our set of layers L,

$$\rho_{ij} = \frac{\sum_{\forall l \in L} \delta_{ij}^{l}}{|L|},$$

with:

$$\delta_{ij}^{l} = \begin{cases} 1 & \text{if } ij \in l, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, ρ_{ij} is the average number of layers in which i is connected to j.

If we calculate the average ρ_{ij} value for all the edges in the issue–issue network shown in Fig. 3, we obtain a value of 0.035, meaning that the average edge appears in only 3.5% of all layers. This seems to suggest that the network is characterized by a certain level of complementarity. The intuition is confirmed if we look at the edges with the highest value of redundancy. The top eight edges, the ones present in 13 or more layers, are either synonyms or antonyms frequently used together: Malnutrition and Nutrition, Communicable Diseases and Non-communicable Disease, AIDS and HIV, and so on. If we were to collapse these issues as one, the average redundancy would be even lower than 0.035.

Moreover, there are 19 connections in the issue–issue network as shown in Fig. 3 that do not appear in a single layer. This means that it is the aggregation of correlated discourses that make these connections emerge, even if there is not a single organization for which the two issues are co-mentioned a significant amount of times. Table A4 reports the full list.

Another multilayer network analysis focuses on estimating the similarities between each layer and the whole structure. We can calculate two related, but complementary measures: Alignment and Impact. Alignment tells us how much a layer is similar to the whole network. To estimate Alignment, we calculate the correlation coefficient between the global issue–issue *WS* values and the ones coming from a single layer. A high Alignment value means that the connection strength between two issues in the eyes of an organization is correlated with the strength in the overall network. In other words, the organization agrees with the community about the relation between the two issues.

On the other hand, Impact tells us how much of the global structure is due to the organization itself. We define Impact as the normalized difference between Alignment and Alignment$*$. Alignment$*$ is the correlation between the *WS* values of the layer and the ones coming from the global network minus the layer itself.[e] A high Impact value means

that, if we were to remove the layer from the global network, the network would change into something rather different than the layer itself. Impact equal to zero would imply that removing the layer from the network would not change the distribution of the *WS* values at all.

Figure 7 (right) shows the relationship between Alignment and Impact across all the organizations. We divide the space in four quadrants by slicing each axis with its median value. We name each quadrant as follows (where "high" means "higher than median", and "low" means "lower than median"):

- *Leaders*. High Alignment and high Impact—e.g. WHO or UNICEF—mean that the organization agrees with the overall *WS* distribution of the aid community and its removal would cause a shift in this distribution. These are organizations which set the discourse of the health community.
- *Followers*. High Alignment and low Impact—e.g. NORAD or DCD—mean that the organization agrees with the overall *WS* distribution of the aid community, but its removal would not change this distribution. These are organizations which follow the discourse of the health community.
- *Explorers*. Low Alignment and high Impact—e.g. World Bank or IMF—mean that the organization disagrees with the overall *WS* distribution of the aid community, but its removal would cause a shift in this distribution. These are organizations which seem to explore different ways to provide health aid and that have enough power to shift the discourse of the community.
- *Strugglers*. Low Alignment and low Impact—e.g. IsDB or UNPBF—mean that the organization disagrees with the overall *WS* distribution of the aid community, and its removal would not change this distribution. These are organizations which are not following the mainstream of the health aid community, but do not have enough power to shift the discourse.

Unsurprisingly WHO, the main actor in health, is in the leader quadrant: its direction emerges as a defining force of the global health arena. Many organizations score a higher impact than WHO, for instance UNICEF and Doctors without Borders, as they likely use their influence in a more visible way for the population at large. The position of World Bank as explorer also makes sense: the World Bank is a large organization which can coordinate a significant human and economic effort, however it is not traditionally thought of as a health provider, and thus has a lower alignment level with the community.

3.4 Case study: World Bank

We now focus our attention on a single organization: the World Bank. The aim of this section is to provide some examples of organization-centric analyses. We focus on the World Bank because it is one of the largest organizations in international aid. Among its many commitments—helping nations to build more equitable societies and to improve fiscal performance and country competitiveness—the World Bank Group (WBG) also works in the health sector. To support countries in reaching the goal of UHC by 2030 (all people and communities can use the promotive, preventive, curative, rehabilitative and palliative health services they need, of sufficient quality to be effective, while also ensuring that the use of these services does not expose the user to financial hardship), the Bank provides financing, state-of-the-art analysis, and policy advice to improve service delivery and expand access to quality, affordable health care. During the period from fiscal year 2000

to 2016, the World Bank invested US$35 billion in the Health, Nutrition and Population (HNP) thematic areas. Over this period, the average annual lending increased significantly from US$1.3 billion to US$2.4 billion. The Bank currently manages an active HNP portfolio of $11.9 billion.

The WBG has focused its health sector investments and research in areas that are especially vital to helping countries achieve UHC by 2030, working closely with donors, development partners, governments, and the private sector. Some of these focus areas include ending preventable maternal and child mortality; reducing stunting and improved nutrition for infants and children; strengthening health systems and health financing; ensuring pandemic preparedness and response; promoting sexual and reproductive health and rights; and the prevention and treatment of communicable diseases.

In light of this, the World Bank is among the main actors in this area, and one of considerable impact—as highlighted in Fig. 7 (right). Moreover, we have access to additional data sources that make the analysis more substantive.

The first question we want to address is whether the connections we extracted from Web patterns reflect the actual activities of an organization. We construct an index measuring the Involvement of the World Bank with a given country. This index is formulated as follows:

$$\mathit{Involvement} = \log P + \log I + \log GP,$$

where:

- P is the number of World Bank financing projects supporting health services approved in a given country over the 2005–2016 period;
- I is the number of World Bank non-financing projects—or advisory Services and analytics supporting health services—approved by the World Bank;
- GP is the number of World Bank operational units delivering the support to health services in a given country over the 2005–2016 period.

We can correlate World Bank's Involvement with *WS*. Figure 8 shows their direct relationship. The Spearman rank correlation between these two measures is 0.6, with a p-value ~ 0.01. Obtaining an almost significant p-value with only 17 observations is remarkable, even more so considering that there is one clear outlier (Romania).

This relationship might be driven by external factors. A country might be mentioned more overall because it is under the spotlight, or because its development level requires for more attention. Here we propose two controls for addressing this objection. The popularity of a country might be estimated looking at the page views of its page in Wikipedia.[f]

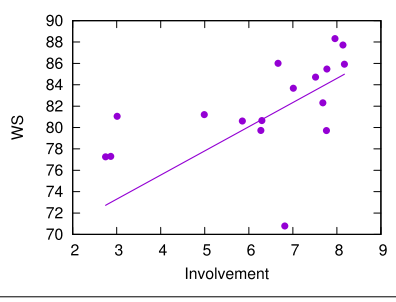


Figure 8 The direct relationship between World Bank's broad support (*x*-axis) and *WS* (*y*-axis). Each observation is a country

Table 6 The direct relationship between World Bank's broad support and *WS*, controlling for a country's popularity and its GDP PPP per capita

	Dependent variable:			
	Involvement			
	(1)	(2)	(3)	(4)
WS	0.126**	0.076*	0.157***	0.103*
	(0.052)	(0.039)	(0.040)	(0.051)
log(wiki.pageviews)		1.695***		1.187
		(0.427)		(0.752)
log(gdp.pc)			0.623***	0.243
			(0.178)	(0.294)
Constant	−3.488	−20.343***	−11.518**	−18.422***
	(4.225)	(5.203)	(3.932)	(5.753)
Observations	17	17	17	17
R^2	0.284	0.663	0.618	0.680
Adjusted R^2	0.237	0.615	0.564	0.606
Residual std. error	0.910	0.646	0.688	0.654
F statistic	5.958**	13.770***	11.344***	9.199***

Note: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Its development level is instead approximated by its GDP per capita PPP. Table 6 shows these controls. Remarkably for such a small number of observations, *WS* is still significant.

To conclude this section, we perform an analysis borrowing tools from network epidemics [19]. Real communication channels between organization do not usually take place on the Web, and thus are outside the data we analyze. However, we can interpret the organization-organization citation network as a trace of *past* communication: if two organizations are talking about each other and they link each other, it means that they are aware of their actions in the field, and are paying attention at each other discourse. In this light, the citation network can be used as a forensic tool to analyze influence after it took place. Note that, in the original network, we establish a directed edge going from organization i to organization j if i cites j. Here, we are interested in knowing that j influenced i, so we reverse the direction of all edges, transforming the citation network into an influence network.

Susceptible–Infected (SI) models have been defined for network data [26]. In these models, nodes are assigned to one of two classes: Susceptible if the node can be infected with a disease, and Infected if the infection happened. In the most simple model—which we use here—a parameter β is specified: if more than a β fraction of the incoming edge weight of a Susceptible node comes from Infected nodes, then the node turns from Susceptible to Infected. These models have been successfully used to track the spread of information in social networks [27].

Here we assume that at time step zero only the World Bank is "infected" with a message it wants to spread. We run an SI model, recording at each time step the fraction of nodes that are part of the Infected pool. Here β represents the share of "infected" cited pages by an organization i necessary to infect i itself. If $\beta = 0.1$, to infect i we require that at least 10% of the pages cited by i have to come from infected organizations. In practice, higher βs mean that we require more citations between organizations A and B to say that A has influenced B in the past.

If we assume that all influence connections in the network can be used, the World Bank is able to infect almost the entire network (∼94% of the nodes). The leftover nodes have

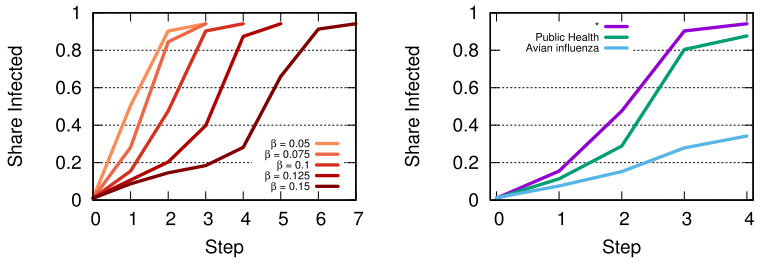


Figure 9 Different infection dynamics in the organization-organization citation network. The lines end when no further infections are recorded in the model. (Left) Infection dynamics considering all edges for different values of the infection parameter β. (Right) Infection dynamics for $\beta = 0.1$ for different keywords: overall (purple), Public Health (green), Avian Influenza (blue)

no sufficiently strong incoming connections, and thus cannot be influenced. This is tested across a variety of β values. Figure 9 (left) depicts the result of these simulations. The higher the β parameter the harder it is for an infection to spread. However, in the case of the World Bank this only affects the speed of propagation of the information (from three to seven steps), not the final coverage. This implies that, according to this model, if the World Bank sends a message to the international aid community, likely 94% of organizations will receive it eventually, assuming that the real unobservable epidemics parameter β is equal to or lower than 0.15.

However, not all messages are equal. The content of the message likely influences its chances to be passed or not. We can simulate also this case, by creating a multilayer view of the influence network, where each layer only contains citations made from pages containing only a specific keyword. We can run the SI model using exclusively edges coming from a single layer, which will now inform us about the power of the World Bank to influence organizations exclusively about a specific issue.

Figure 9 (right) depicts the result of these simulations. We also report the result of the simulation using all the layers, for reference. We can see that there are significant differences between issues. The World Bank is able to reach the vast majority of nodes in some cases, for instance when talking about Public Health (the final infected share of nodes is $\sim$88%). However, in the case of Avian Influenza the message finds a bottleneck in the multilayer network, and only reaches a third of the network.

How does the World Bank compare to other organizations? Here we choose three comparisons: WHO, UNICEF and USAIDS. We run the SI model using all connections from all keywords, and fixing β once again to 0.1. Figure 10 (left) depicts the result. We can see that the World Bank is noticeably slower than WHO, which reaches saturation faster. However, the World Bank outperforms UNICEF. When considering all keywords at the same time, the messages coming from USAID are dwarfed and reach a negligible portion of the network.

However, as pointed out before, these diffusion patterns are highly dependent on which keyword we are focusing. Figure 10 (right) depicts the information spreading results when focusing on a specific one: "Nurse". In this case, the four organizations are hardly distinguishable, with USAID having an influence potential on par with the World Bank and WHO.

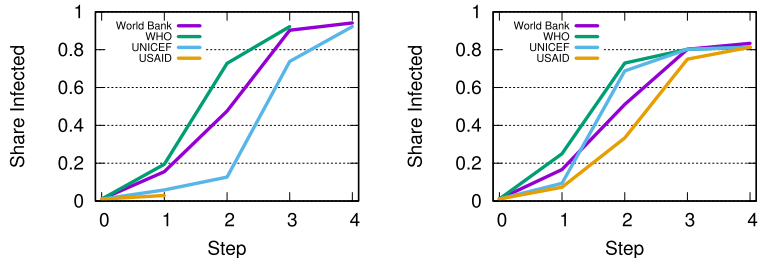
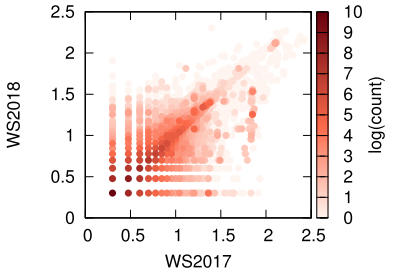


Figure 10 Different infection dynamics in the organization-organization citation network. The lines end when no further infections are recorded in the model. (Left) Infection dynamics for $\beta = 0.1$ for different organizations, aggregated to all keywords. (Right) Same as left, but only considering connections for the topic "Nurse"

Figure 11 The relationship between the *WS* when calculated using data collected in the February 2017 crawl (*x*-axis) and the January 2018 crawl (*y*-axis). Each data point is colored according to how many edges had the particular weight combination (log scale)



3.5 Stability

The Web is notoriously dynamic: there is no guarantee that a webpage you can visit now is going to be online tomorrow. In the face of such dynamism, one could argue that any snapshot gathered from the Web could be noisy to the point of not being a good representation of the system.

Our original crawl was performed in February 2017. Eleven months after, on January 2018, we perform a second crawl of a selected number of websites to test the stability of our results. We selected roughly half of the websites and apply our data cleaning pipeline. The question is: is the 2017 *WS* measure a good predictor of 2018 *WS*? We answer by doing two analysis.

First, we perform a simple regression analysis. Regressing 2017 *WS* against 2018 *WS* yields an $R^2 \sim 74.7\%$. This means that the vast majority of the variance in the 2018 measure can be predicted using 2017 data. This result reassures us that, overall, the networks we build are stable over time. A perfect correlation would have been as problematic as a low one: it would imply that the international aid community does not change which would be a puzzling property for such a complex system. Since the correlation is not perfect, we are reassured that we are able to capture the evolution of the system.

We also provide a visual representation of the relationship between the 2017 and 2018 versions of *WS*, in Fig. 11. Most of the data points lie in the diagonal, which means that most edges maintained the same weight.

4 Discussion

In this paper we show how one can use web crawling to recreate networks of relations between issues, countries, and organizations working in the international aid community.

We show how the results are meaningful and valid. For instance, there is a strong relationship between the relatedness of World Bank to issues/countries as emerging from the Web data and as estimated by measures of actual resource allocation.

However, there are also several aspects that make the data suboptimal, and a cause for future improvements. In this section, we list the most important. They relate to noise, valence, counter-intuitive results, and cultural bias.

When it comes to noise, our paper is a significant improvement over previous attempts [3]. Noise implies that the number of co-occurrences of two terms is improperly measured, because of different reasons. Previous work relied on the index created by Google. Such a huge index provides less precise hit counts for specific queries [13]. By directly crawling the information, our paper reduces the measurement error. We also employ more sophisticated noise-correction when it comes to network cleaning [20]. Notwithstanding these improvements, web data is still affected by large potential for noise (e.g. a connection drop might prevent the crawler to access entire sections of a website). As a consequence, there is room for the improvement of the data collection and cleaning phase.

Using co-occurrences in web pages to connect two entities ignores the rest of the content of the page. The article in which two organizations were mentioned could be related to positive or negative views of such relationship. At the moment, we ignore the valence of this connection. Thus this study limits itself to saying that the two entities are related "somehow". It is conceivable to extend this work by having both positive and negative connections, e.g. by performing sentiment analysis on the content of each web page [28].

This issue can lead to some counter-intuitive results. For instance, if we create a network with relations between organizations in a specific country, we might include in such network also agencies that do not operate in that specific world area. This will happen when another organization publishes a page containing both the country name and the agency name. Such decision might have different reasons, for instance it could be an attempt to raise awareness of such missed connection. Currently, we keep these connections as traces of a signal connecting the agency to the place—even if not physically, at least in the intentions of another organization.

Finally, we focus exclusively on web pages written in the English language. Even if we include non-English pages, our data collection strategy will filter them out, since our issue list is specified in English. Having an unambiguous translation of all issues in many languages is not a trivial operation. Since the vast majority of international aid organizations have an English website, we conclude that this issue has a minor impact on the results of this paper.

5 Conclusion

In this paper we describe how to perform web crawling to create a network view of the international aid community. We focus in particular on organizations and issues related to health. We validate our results by showing the correspondence between the intensity of organization-country and organization-issue web relationships and the actual amount of resources that the organization invests in the country/issue. By analyzing the collected network, we can describe some of the salient characteristic of the system. Organizations tends to be co-cited in groups, according to their type: multilateral, bilateral, private, etc. Topics can be clustered in a taxonomy of areas of interest. Finally, we show that the citation network has a single connected component, meaning that in theory messages originating

from one organization can reach the entirety of the network. However, in practice, for different issues there are different bottlenecks that might obstruct the message-passing.

This paper can be extended in several ways. First, we can expand the set of organizations, countries, and topics, to have a more representative view of international aid. Second, we can extend the languages considered in the web crawl, by providing unambiguous translations of the issues on which we focus. Finally, the paper only presented a subset of the possible analysis that can be done using the collected data. Different views of the multilayer networks can be further explored to improve our description of the international aid complex system.

Appendix: Entity lists

Table A1 The list of countries on which this study focuses

Country	Alias
Liberia	
Sierra Leone	
Nigeria	
Indonesia	
Philippines	
Vietnam	
Romania	
Kyrgyz Republic	Kyrgyzstan
Turkey	
Peru	
Brazil	
Haiti	
Bangladesh	
Afghanistan	
Sri Lanka	
Egypt	
Jordan	

Table A2 The list of organizations on which this study focuses

Full Name	Acronym	URL	Aliases
Austrian Development Agency	ADA	http://www.entwicklung.at/en/	
Belgian Development Agency	BTC	https://www.btcctb.org/	
Czech Development Agency		http://www.czechaid.cz/en/	
Danish International Development Agency	DANIDA	http://um.dk/en/danida-en/	
Department for International Development, UK	DfID	https://www.gov.uk/government/organisations/department-for-international-development	DFID
Department of Development Cooperation, Poland	DCD	https://www.polskapomoc.gov.pl/	Polish aid
Department of Foreign Affairs and Trade, Australia		http://dfat.gov.au/aid/	AusAid; Aus Aid
Department of Foreign Affairs, Canada		http://www.international.gc.ca/development-developpement/index.aspx?lang=eng	CIDA
Deutsche Gesellschaft für Internationale Zusammenarbeit	GIZ	https://www.giz.de/en/	GTZ

Table A2 (*Continued*)

Full Name	Acronym	URL	Aliases
Economic Development Cooperation Fund, Korea	EDCF	http://211.171.208.43/edcfeng/index.jsp	
French Development Agency	AFD	http://www.afd.fr/lang/en/	
Icelandic International Development Agency	ICEIDA	http://www.iceida.is/english/	
Irish Aid		https://www.irishaid.ie/	
Italian Development Cooperation		http://openaid.esteri.it/en/	
Japan International Cooperation Agency	JICA	https://www.jica.go.jp/english/	
Korea International Cooperation Agency	KOICA	http://www.koica.go.kr/english/main.html	
Kreditanstalt für Wiederaufbau	KfW	https://www.kfw-entwicklungsbank.de/International-financing/KfW-Entwicklungsbank/	KFW
Lux-Development	LuxDev	https://luxdev.lu/en/	Lux Development
Millennium Challenge Corporation	MCC	https://www.mcc.gov/	
Ministry for Foreign Affairs of Finland	FINNIDA	http://formin.finland.fi/	
Ministry of Foreign Affairs of the Netherlands		https://www.government.nl/topics/development-cooperation/	
Ministry of Foreign Affairs, Slovenia		http://www.mzz.gov.si/en/	
New Zealand Foreign Affairs & Trade		https://www.mfat.govt.nz/en/aid-and-development/	
Norwegian Agency for Development Cooperation	NORAD	https://www.norad.no/en/front/	
Slovak Agency for International Development Cooperation	SAIDC	http://www.slovakaid.sk/en/	SlovakAid; Slovak Aid
Spanish Agency for International Development Cooperation	AECID	http://www.aecid.es/EN/	
Swedish International Development Cooperation Agency	SIDA	http://www.sida.se/English/	
Swiss Agency for Development and Cooperation	SDC	https://www.eda.admin.ch/sdc/	
The U.S. President's Emergency Plan for AIDS Relief	PEPFAR	http://www.pepfar.gov/	
United States Agency for International Development	USAID	https://www.usaid.gov/	
Abu Dhabi Fund for Development	ADFD	https://www.adfd.ae/english/	
Federal Agency for the Commonwealth of Independent States, Compatriots Living Abroad and International Humanitarian Cooperation	Rossotrudnichestvo	http://rs.gov.ru/en/	
Kuwait Fund For Arab Economic Development	KFAED	https://www.kuwait-fund.org/en/web/kfund/	Kuwait Fund
Malaysian Technical Cooperation Programme	MTCP	https://mtcp.kln.gov.my/	
National Health and Family Planning Commision of the PRC		http://en.nhfpc.gov.cn/	
Saudi Fund for Development	SFD	http://www.sfd.gov.sa/	
Singapore Cooperation Programme	SCP	https://www.scp.gov.sg/content/scp/	Singapore Cooperation Program
Thailand International Cooperation Agency	TICA	http://www.tica.thaigov.net/main/en/	
African Development Bank	AfDB	http://www.afdb.org/en/	AFDB
Arab Bank for Economic Development in Africa	BADEA	http://www.badea.org/	

Table A2 (*Continued*)

Full Name	Acronym	URL	Aliases
Arab Fund for Economic & Social Development	AFESD	http://www.arabfund.org/	Arab Fund
Asian Development Bank	ADB	https://www.adb.org/	
Asian Infrastructure Investment Bank	AIIB	http://www.aiib.org/	
Caribbean Development Bank	CDB	http://www.caribank.org/	
Council of Europe Development Bank	CEB	http://www.coebank.org/en/	
Development Bank of Latin America	CAF	https://www.caf.com/en/	
European Bank for Reconstruction and Development	EBRD	http://www.ebrd.com/	
European Investment Bank	EIB	http://www.eib.org/	
European Union	EU	http://ec.europa.eu/europeaid/	European Commission
Inter-American Development Bank	IDB	http://www.iadb.org/en/	IADB; BID; Banco Inter-americano de Desarollo
International Finance Corporation	IFC	https://www.ifc.org/	
International Fund for Agricultural Development	IFAD	https://www.ifad.org/	
International Monetary Fund	IMF	http://www.imf.org/	
Islamic Development Bank	IsDB	http://www.isdb-pilot.org/	ISDB
Multilateral Investment Guarantee Agency	MIGA	https://www.miga.org/	
New Development Bank	NDB	http://ndb.int/	BRICS Bank
Nordic Development Fund	NDF	http://www.ndf.fi/	
OPEC Fund for International Development	OFID	http://www.ofid.org/	
World Bank	WB	http://www.worldbank.org/	
Food and Agriculture Organization	FAO	http://www.fao.org/home/en/	
International Labour Organization	ILO	http://www.ilo.org/	
Joint United Nations Programme on HIV/AIDS	UNAIDS	http://www.unaids.org/	
United Nations Children's Fund	UNICEF	http://www.unicef.org/	
United Nations Development Programme	UNDP	http://www.undp.org/	
United Nations Foundation	UNF	http://www.unfoundation.org/	
United Nations High Commissioner for Refugees	UNHCR	http://www.unhcr.org/	
United Nations Peacebuilding Fund	UNPBF	http://www.unpbf.org/	
United Nations Population Fund	UNFPA	http://www.unfpa.org/	
United Nations Relief and Works Agency for Palestine Refugees in the Near East	UNRWA	http://www.unrwa.org/	
World Food Programme	WFP	http://www.wfp.org/	
World Health Organization	WHO	http://www.who.int/en/; http://www.euro.who.int/en/; http://www.afro.who.int/; http://www.searo.who.int/en/; http://www.emro.who.int/	
Pan-American Health Organization	PAHO	http://www.paho.org/hq/	OPS; Organización Panamericana de la Salud
African Program for Onchocerciasis Control	APOC	http://www.who.int/apoc/en/	ONCHO
Global Alliance for Improved Nutrition	GAIN	http://www.gainhealth.org/	
Global Alliance for Vaccines and Immunization	GAVI	http://www.gavi.org/	Gavi; the Vaccine Alliance

Table A2 (*Continued*)

Full Name	Acronym	URL	Aliases
Global Financing Facility	GFF	http://globalfinancingfacility.org/	
Global Health Investment Fund	GHIF	http://www.ghif.com/	
Harnessing Non-State Sector for Better Health for the Poor	HANSHEP	http://www.hanshep.org/	
Health Results Innovation Trust Fund	HRITF	https://www.rbfhealth.org/	RBF
International AIDS Vaccine Initiative	IAVI	http://www.iavi.org/	
International Health Partnership and related Initiatives	IHP	http://www.internationalhealthpartnership.net/	
Medicines for Malaria Venture	MMV	http://www.mmv.org/	
PAHO Revolving Fund		[Subset of PAHO's webpages]	
Partnership for Maternal, Newborn and Child Health	PMNCH	http://www.who.int/pmnch/en/	
Power of Nutrition		http://www.powerofnutrition.org/	
Roll Back Malaria	RBM	http://www.rollbackmalaria.org/	
Scaling up Nutrition	SUN	http://scalingupnutrition.org/	
Special Program for Research and Training in Tropical Diseases	TDR	http://www.who.int/tdr/en/	
Stop TB	Stop TB	http://www.stoptb.org/	
The Global Fund to Fight AIDS, Tuberculosis and Malaria	GFATM	http://www.theglobalfund.org/en/	The Global Fund
UNITAID	UNITAID	http://www.unitaid.eu/en/	
Bill & Melinda Gates Foundation		http://www.gatesfoundation.org/	Gates Foundation
MasterCard Foundation		http://www.mastercardfdn.org/	
Rockefeller Foundation		https://www.rockefellerfoundation.org/	
CARE International	CARE	http://www.care-international.org/	CARE
International Federation of Red Cross and Red Crescent Societies	IFRC	http://www.ifrc.org/	International Red Cross; International Red Crescent
Médecins Sans Frontières	MSF	http://www.msf.org/; http://www.doctorswithoutborders.org/	Doctors Without Borders
Oxfam International	OXFAM	https://www.oxfam.org/	Oxfam
Plan International		https://plan-international.org/	Plan
Save the Children International		https://www.savethechildren.net/	Save the Children
World Vision International	WVI	http://www.wvi.org/	World Vision
Center for Global Development	CGD	http://www.cgdev.org/	
International Planned Parenthood Federation	IPPF	http://www.ippf.org/	

Table A3 The list of issues on which this study focuses

AIDS, Avian influenza, Child Health, Communicable Diseases, DAH, Development Assistance for Health, Disease Control, Disease Outbreak, Disease Prevention, Ebola, Family Planning, Global Health, Health Care, Health Center, Health Centre, Health Clinic, Health Economics, Health Facility, Health Financing, Health Infrastructure, Health Insurance, Health Intervention, Health Market, Health Outcome, Health Policies, Health Policy, Health Portfolio, Health Reform, Health Sector, Health Security, Health Service, Health stewardship, Health Strategy, Health System, HIV, Hospital, Immunization, Infant Survival, Infectious Disease, International Health, Malaria, Malnutrition, Maternal Health, Medical, Medicine, Mental Health, Ministry of Health, Mortality Rate, NCDs, Neglected disease, Non-communicable Disease, Nurse, Nutrition, Occupational Health, Pandemic, Pharmaceutical, Polio, Primary Care, Primary Health, Public Health, Reproductive Health, Secondary Care, Stunting, TB, Tertiary Care, Tropical Medicine, Tuberculosis, UHC, Universal Health, Vaccination, Vaccine.

Table A4 The 19 connections in the issue–issue network that are not significantly connected in any of the crawled websites, but are significantly related in the aid community as a whole

Issue 1	Issue 2
Development Assistance for Health	Health stewardship
Health stewardship	Secondary Care
Health Portfolio	Health stewardship
Disease Outbreak	Health Security
Avian influenza	Disease Prevention
Neglected disease	Tropical Medicine
Health Insurance	Health Reform
Health Reform	Universal Health
Health Outcome	UHC
Health Financing	Health Outcome
Mortality Rate	Stunting
AIDS	Neglected disease
Disease Outbreak	Infectious Disease
HIV	Neglected disease
Health Outcome	Universal Health
Child Health	Stunting
Health Sector	Universal Health
Health Facility	Health Service
Ebola	Global Health

Acknowledgements
We thank Cristano Buizza for the help with data on the intensity of the World Bank's support to health services. We thank Jos Vaessen for valuable comments.

Funding
Michele Coscia has been partly supported by FNRS, grant #24927961.

Availability of data and materials
The input and output data, and the code of the web crawler are freely available at http://www.michelecoscia.com/?page_id=1307.

Competing interests
Support from the Independence Evaluation Group of The World Bank Group is acknowledged. The organization did not play any role in shaping the analysis and the draft of this paper. The authors declare no conflict of interest.

Authors' contributions
Michele Coscia developed the methods and performed the experiments. Antonio Giuffrida managed the project. Antonio Giuffrida, Katsumasa Hamaguchi, and Maria Elena Pinglo provided the input data. All authors designed the experiments. All authors wrote, reviewed, and approved the manuscript.

Author details
[1]Naxys Department of Mathematics, University of Namur, Namur, Belgium. [2]Center for International Development, Harvard University, Cambridge, USA. [3]Independent Evaluation Group, World Bank, Washington, USA.

Endnotes
[a] http://www.michelecoscia.com/?page_id=1307
[b] https://scrapy.org/
[c] http://phantomjs.org/
[d] More precisely, if n organizations were mentioned, the page would be counted n times. One can easily see that $WS_i \leq \sum_{\forall o \in O} WS_{oi}$.
[e] Note that this is very similar to calculating the correlation between the global network and the global network minus the layer, but this latter measure is harder to interpret.
[f] A related, but less reliable measure would be the number of articles in online newspapers. Unfortunately, the counts in Google News are less reliable than page counts in Wikipedia. Moreover, there are ambiguities to consider, e.g. "Jordan" is a popular name in English speaking countries.

Publisher's Note

Received: 31 October 2017 Accepted: 30 April 2018 Published online: 08 May 2018

References

1. Hoffman SJ, Cole CB, Pearcey M (2015) Mapping global health architecture to inform the future. Chatham House, London
2. Wu B (2016) Global Dynamics: Approaches from Complexity Science, pp 231–256
3. Coscia M, Hausmann R, Hidalgo CA (2013) J Glob Dev 3(2):1
4. Albert R, Barabási AL (2002) Rev Mod Phys 74(1):47
5. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Phys Rep 424(4):175
6. Newman ME (2003) SIAM Rev 45(2):167
7. Lee SH, Kim PJ, Ahn YY, Jeong H (2010) PLoS ONE 5(7):e11233
8. Coscia M, Rios V (2012) Proceedings of the 21st ACM international conference on information and knowledge management. ACM, New York, pp 1412–1421
9. Popović M, Štefančić H, Sluban B, Novak PK, Grčar M, Mozetič I, Puliga M, Zlatić V (2014) PLoS ONE 9(12):e99515
10. Liu Y, Wang F, Kang C, Gao Y, Lu Y (2014) Trans GIS 18(1):89
11. Lim YS, Park HW (2013) New Media Soc 15(1):93
12. Lang PB, Gouveia FC, Leta J (2013) PLoS ONE 8(8):e71415
13. Funahashi T, Yamana H (2010) Current trends in web engineering, pp 114–125
14. Jacsó P (2006) Online Inf Rev 30(2):188
15. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) J Complex Netw 2(3):203
16. Berlingerio M, Coscia M, Giannotti F, Monreale A, Pedreschi D (2011) International conference on advances in social networks analysis and mining (ASONAM). IEEE, New York, pp 485–489
17. Berlingerio M, Coscia M, Giannotti F (2011) Proceedings of the 20th ACM international conference on information and knowledge management. ACM, New York, pp 2181–2184
18. McPherson M, Smith-Lovin L, Cook JM (2001) Annu Rev Sociol 27(1):415
19. Pastor-Satorras R, Vespignani A (2001) Phys Rev Lett 86(14):3200
20. Coscia M, Neffke F (2017) Preprint. arXiv:1701.07336
21. Grady D, Thiemann C, Brockmann D (2012) Nat Commun 3(3):864
22. Newman ME (2006) Proc Natl Acad Sci USA 103(23):8577
23. Rosvall M, Bergstrom CT (2008) Proc Natl Acad Sci USA 105(4):1118
24. Barthelemy M (2004) Eur Phys J B, Condens Matter Complex Syst 38(2):163
25. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Comput Netw 33(1):309
26. Karsai M, Kivelä M, Pan RK, Kaski K, Kertész J, Barabási AL, Saramäki J (2011) Phys Rev E 83(2):025102
27. Kwak H, Lee C, Park H, Moon S (2010) Proceedings of the 19th international conference on World Wide Web. ACM, New York, pp 591–600
28. Pang B, Lee L et al (2008) Found Trends Inf Retr 2(1–2):1