MetaMetaZipf. What do analyses of city size distributions have in common?

Clementine Cottineau, CNRS

1. Introduction

The parallel development of more data accessible at city level on the one hand and of non linear regularities being found as a marker of complexity (in networks, in organisms, in cities) have produced a regain of interest in the study of city size distributions in recent years, and especially in the discussion of Zipf's law for cities. Of particular importance in these debates are the definition of the objects studied (i.e. the limits, thresholds and components of cities which affect the population included or not), the model to summarize the distribution (power-law, lognormal, polynomial) and its fit to the data (fitting procedure, value of the power exponent, uncertainty). It is usually agreed that city populations follow a heavy-tail population in most countries and at most time periods, although the precise form of the distribution and the estimation of its main parameters tends to vary. The universality claim of Zipf's law (1949) can thus be accepted with respect to the general trend, but has to be rejected in its strictest form (i.e. a power law of exponent -1 between city sizes and their ranks by size). Previous meta-analyses of studies providing an empirical estimation of Zipf's exponent have shown indeed that on average, empirical estimations deviate from the strict value of -1 (Nitsch, 2005; Cottineau, 2017). A share of such deviations can be attributed differences in the technical specifications of the studies (its number of estimates, range of countries and periods analysed) and of the empirical estimation (delineation of cities, thresholds, estimation procedure, etc.). A smaller share can be attributed to territorial characteristics of the city system (its phase and level of urbanisation) and no share can be attributed to major planning actions. However, empirical deviations to Zipf's law remain for the most part unexplained (Cottineau, 2017) or unexplored. For instance, publication biases as well as differences in reference frameworks and disciplinary traditions might generate systematic differences in measuring and reporting of empirical distributions of city sizes which are unobservable with a traditional meta analysis. For example, despite addressing the same empirical estimation of Zipf’s law (same country, same set of city, same date, same estimation method), there can be strong differences in the way the papers from the meta analysis frame, exploit and report on this result, depending on the aim of their research (such as “proving that Zipf's law is a universal feature of urban system”, “showing that the lognormal form is better suited”, “looking for national differences in urban hierarchy”, etc.). The empirical results of such study could then appear clustered by different school of thoughts. The present work therefore goes a step further in the secondary analysis of Zipf's law for cities, by exploiting various network properies desmonstrated by the studies of a meta analysis tehmselves. Building on the open-source corpus of MetaZipf (Cottineau, 2017), which contains 1962 empirical estimations of Zipf's exponent alongside their technical and territorial specifications from 86 studies, it characterises the pairwise simliarities of studies based on their bibliographies, their textual content and their disciplinary exposure. Combined with the pairwise similarities of the study content, it aims to reveal new insights regarding the deviation in their published results. We find evidence that pairs of articles with similar wording and similar bibliographies tend to report similar average values of estimates. Similar wording also correlates positively with a similarity in the level of dispersion of values reported.

2. Why a meta-meta-analysis?

Meta-analyses are important tools to summarize and reflect on the collective production of an established field of enquiry, especially when it produces quantitative estimations and prediction statements. In that respect, city size distributions and their modelling with power laws dates back more than a century (Auerbach, 2013), and still generates dozens of dedicated articles each year. However, such scientific productions originate from a diversity of displines and research domains

such as economics, geography, statistics, physics, regional science, planning and mathematics. Furthermore, authors of studies included in the Zipf meta-analysis publish in a diversity of journals which all have their different formal (size of text, types of proofs received as valid) and theoretical requirements (references considered as necessary, legitimate, or superfluous). For instance, economics journal will require econometric models with controls and way of presenting results in standardised tables. Physics journals tend to publish short articles with large supplementary materials. "*Planning papers tend to cite eclectically. [...] This will be a feature of social science in general compared with science journals but, within the social sciences, one might expect certain broader applied subjects such as planning to be especially unfocused in the literature they cite. […] Planning papers are also eclectic in the type of references cited reports and plans as well as academic papers and this may lower impact statistics.*" (Webster, 2006, p.488). Journals in geography will tend to favour analyses of spatial variations of a given phenomenon while other subjects will look for its regularity. Could such meta properties of articles dedicated to the empirical estimation of Zipf's law play a role in the definition of the aim of the research, the design of the experiment and eventually the value of the reported results, offering a new angle to explain their difference? The hypothesis leading this research is that it could. Indeed, science is a social practice performed by social actors embedded within institutions, disciplinary frameworks and legacies (Latour, 1986). It would therefore be possible to find that studies written in a similar was, citing similar references and publishing in the same kind of journal would exhibit more similar reported results (controlling for the object of their sutdy, in our case, the similarity of cities, countries and time periods studied) than studies which originate from very different fields, points to very different sets of bibliographic refences and use different scientific language. There is evidence from the MetaZipf corpus that a significant diversity of language, reference framework and disciplinary display exist. In the way articles are written for example, Gabaix & Ibragimov's (2011) article is built like a mathematical demonstration (using terms like "theorem" seven times and "lemma" four times) whereas other articles read more like monographies. With respect to the way the MetaZipf articles cite other works, some articles systematically reference back to Zipf (1949) and Auerbach (1913), whereas others start the debates to where Gabaix (1999) left it. Some articles cite a very large amount to external references (Parr, 1985 or Berry & Okulicz-Kozaryn, 2012) when others do not (such as early articles and short pieces published in physics journals). Finally, the range of journals cited and chosen for publication is broad, and ranges from mainstream economics to specialised geography to statistical physics and beyond. The objective of the present work is to assess whether such diversity is reflected systematically in the variations of results reported and to which extent it can contribute to better understand urban hierarchies around the world (rather than the individuals who study them).

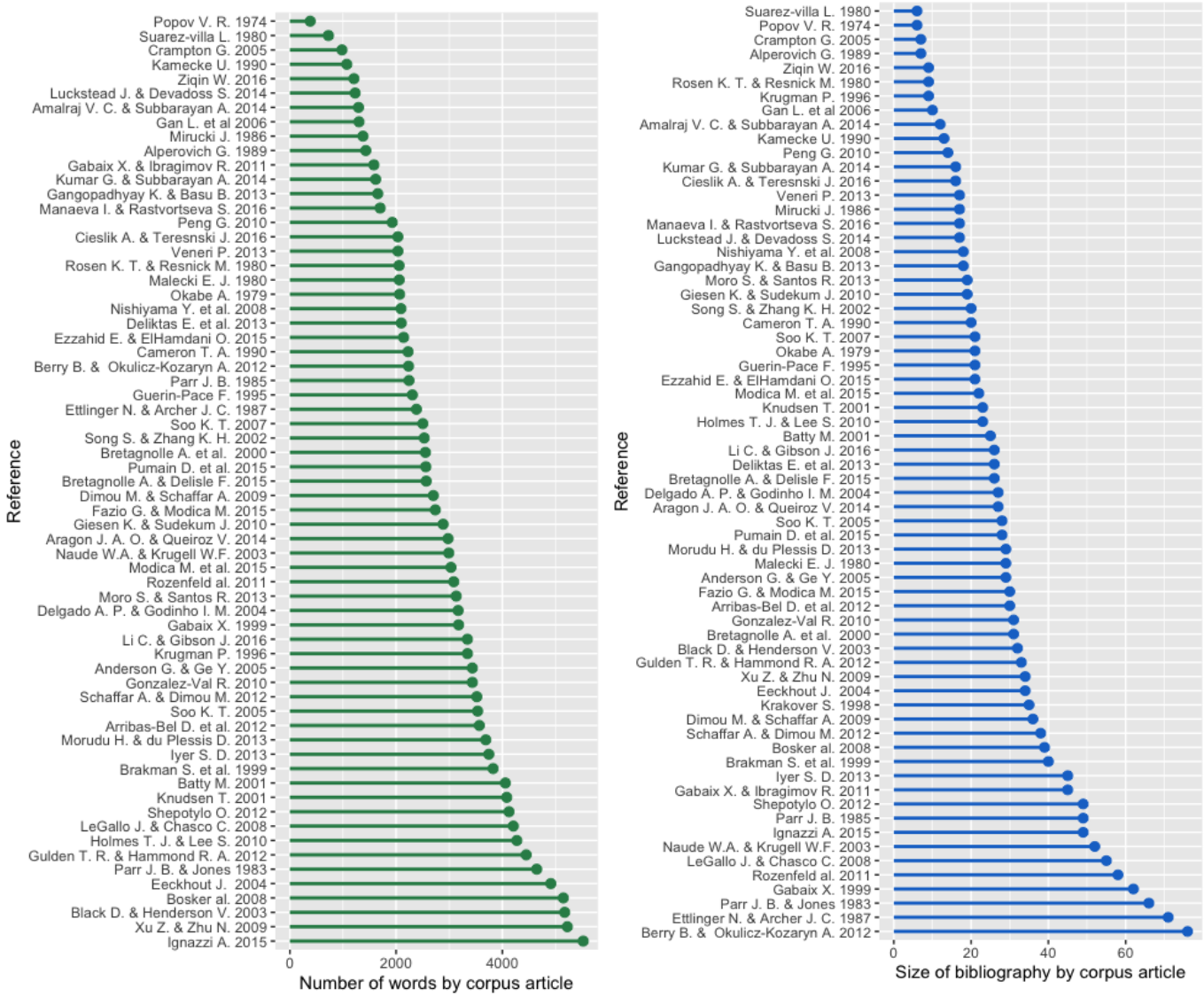
3. Methods and materials

This section details the collection of some meta-meta-data and the strategy used to convert network into pairwise similarity matrices along a number of dimensions. It also presents the model used to regress differences in Zipf estimations by scientific practice, controlling for technical and territorial specifications. The material of the present study consists in a corpus of studies which have published empirical estimations of Zipf's exponent on city population, as of 2017. I make use of the openly available database MetaZipf (https://github.com/ClementineCttn/MetaZipf), which contains 1962 empirical estimations of Zipf's exponent along with their technical and territorial specifications from 86 studies. The 86 studies have been selected to fulfill three criteria: "*they contain at least one estimate of the rank-size exponent based on population ; the regression is made on empirical urban data; the regression model is bivariate (i.e. relating populations and ranks or ranks—1/2, but not to any other instrumental variable).*" (Cottineau, 2017, p. 4). In the present work, only 66 of them fulfilled additional criteria detailed below. This subset of 66 studies are subsequently referred to as "the corpus".

3.1. Collecting full-texts

For an article from the MetaZipf database to be included in "the corpus", it had to available in open-access or accessible with an extensive institutional subscription in a pdf-readable format. Additionally, in order to run a coherent textual analysis, only published journal articles written in English were kept. This excluded texts in other languages and formats, such as books and dissertations. This choice is detrimental to the recognition that science is plural in forms, languages and origins. However, it did not affect the original sample too much, since most references in MetaZipf were already predominantly in English and in an article format. The corpus is thus composed of 66 full-texts of English-written articles. Out of the original document, only the body of the text was retained. This means that titles, affiliations abstracts, keywords, section titles, figures, tables, equations, references, footnotes and line breaks were removed. The remaining text was used for text mining analysis, after a traditional automated treatment (with the R 'tm' package, cf. Feinerer et al., 2019) in order to remove punctuation, numbers and stop-words and transform the remaining word to lower case. Term frequencies were attached to each reference to allow for a study of wording similarity between them.

Figure 1. A (left). Distribution of text size in the corpus (number of non-stop-words). B (right). Distribution of bibliography size in the corpus (number of external references).



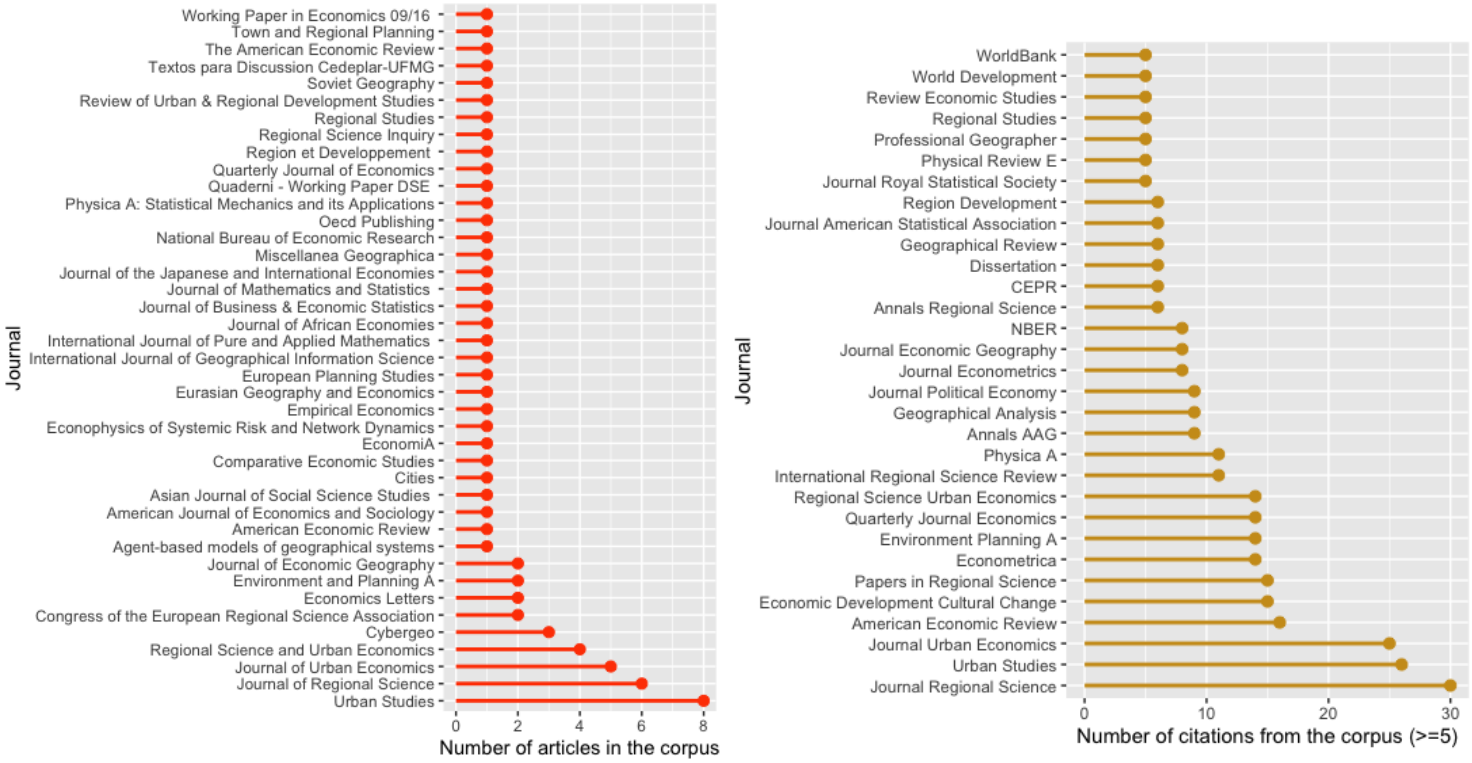
Once this procedure is complete, corpus articles exhibit a continuous array of sizes (figure 1A), from 384 for Popov (1974) to 5522 for Ignazzi (2014). Apart from significantly shorter sizes in

physics articles (around 1600 words on average per corpus article, compared to 3000 on average in economics and 2500 in geography), we could not find any trend by year of publication or else.

3.2. Collecting citations

To explore the citation network of corpus articles, each reference from the 66 english-written articles was recorded and formatted in a way that allows to query the authors' names, the year and journal of publication. The 66 corpus articles generated 304 internal citations (i.e. to other articles included in the corpus) and citations to 1155 distinct external references (including to articles, reports, books or dissertations in various languages) from over 700 different journals or publishing institutions. Corpus articles exhibit once again a disparity of bibliography sizes (figure 1B), from 6 items in Suarez-Villa (1980) and Popov (1974) to 76 in Berry & Okulicz-Kozaryn (2012). Apart from significantly shorter sizes in physics articles (around 15 items on average, compared to 22 on average in economics and 24 in geography and regional science), we could not find any trend by year of publication or else.

Figure 2. A (left). Distribution of corpus articles by journals (and series) publishing them. B (right). Distribution of articles cited by corpus articles by journals (and series) publishing at least 5 of them.

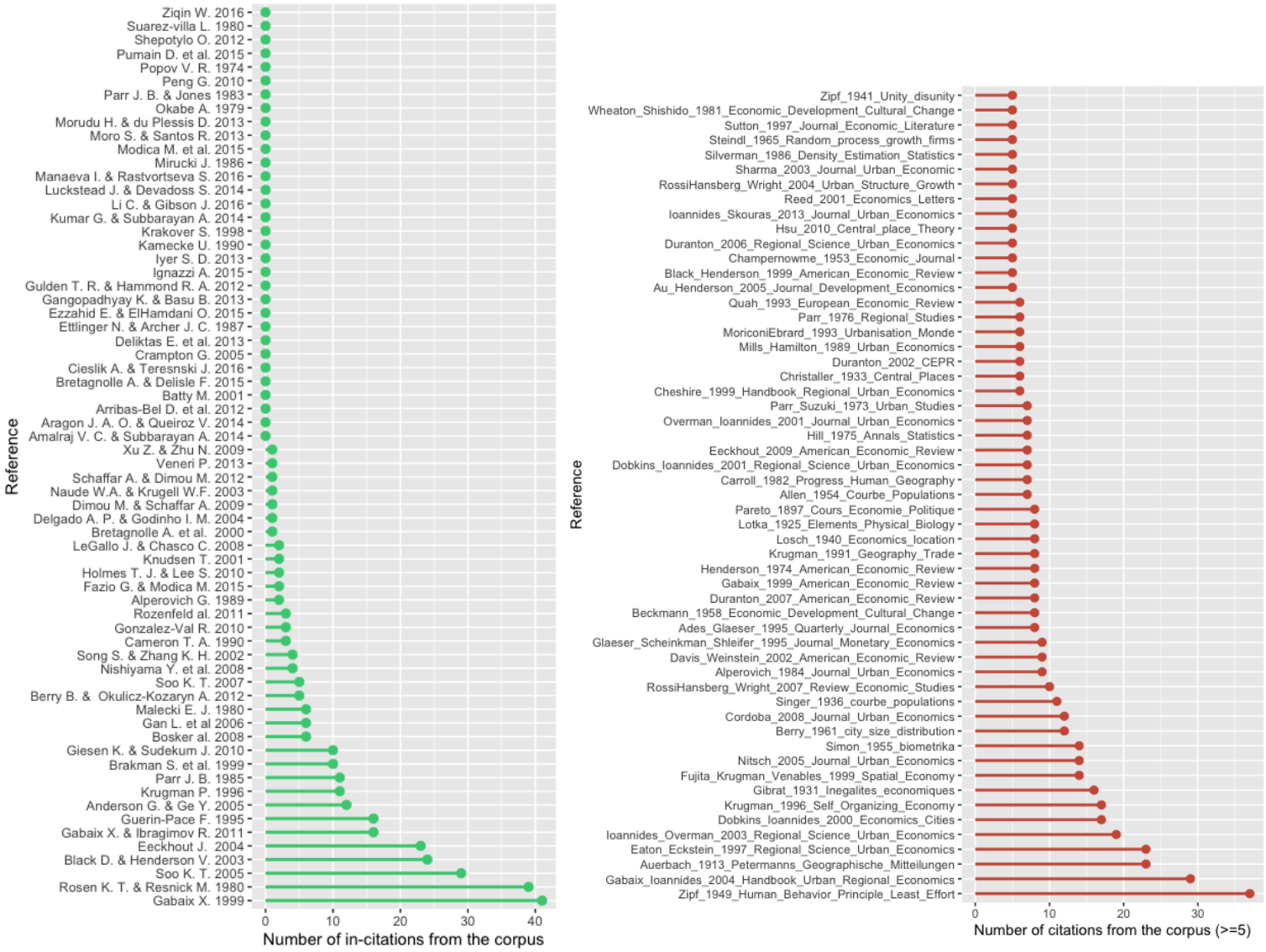


The journals chosen to publish the most corpus articles (figure 2A) coincide with the journals where bibliographical references most frequently come from (figure 2B), i.e. Urban Studies and the Journal of Regional Science, then the Journal of Urban Economics, Regional Science and Urban Economics or the Journal of Economic Geography. The average year of publication in the corpus is 2004, plus or minus on year for articles of different discipline except articles published in physics journals, whose interest in city size distributions and average year of publication is much more recent (2013). By contrast, the average year of publication references in corpus articles 1989.

The most cited external reference is to Zipf himself, with 37 out of the 66 sample articles citing it for its 1949 book on the principle of the least effort and 5 citing it for its 1941 work "National unity and disunity; the nation as a bio-social organism". The papers not citing any of the two Zipf references are frequent among those published at earlier dates (figure S1 in Supplement), and

proportionally more in geography and economics journals where it is considered an evidence, whereas 4 out of 4 articles from physics journals cite Zipf in their paper on city size distributions. It is interesting to note, however, that Zipf's work is not the most cited references in the corpus, as two internal references appear even more frequently (figure 3A): Gabaix's theoretical 1999 paper (41 times out of the 50 other articles published in or after 1999) and Rosen & Resnick's comparative 1980 paper (39 out of 62 possibilities). Externally (figure 3B), the reference to Auerbach's work from 1913 is also in the top 3 of external references, but less prevalent (cited by only 23) and topped by Gabaix & Ioannides's (2004) chapter in the Handbook of Urban and Regional economics with external citations from 29 sample articles.

Figure 3. A (left). Distribution of citations to corpus articles by corpus articles. B (right). Distribution of citations (over 5) to non-corpus articles by corpus articles



The graph on figure 3B shows that many top references externally cited are early classics of urban theory (Christaller, 1933 [6 cites], Losch, 1940 [8]) and statistics (Gibrat, 1931 [16 cites], Simon, 1955 [14], Pareto, 1897 [8], Hill, 1975 [7]). Some highly cited references such as Singer, 1936 [11] or Eaton & Eckstein, 1997 [23], which include empirical estimations of Zipf's exponent, suggest that they could have been included in the corpus. However, the former was not accessible to the author and the latter contains instruments in the regression. It could be considered to include its findings in the future, given its influence on the corpus' reference frameworks.

The most striking feature of this list however is the prominence of post-1995 contributions from three economists in the top cited references (Gabaix, Krugman and Ioannides) compared to earlier works by geographers (like Berry in 1961, Parr since the 1970s, or Moriconi-Ebrard in the early 1990s). As pointed by C. Webster (2006, p. 489-90) in the context of planning journals, "*there is both a publishing and a cognitive limitation on the number of citations included in a paper and this means that the rate of citation growth will be higher, the higher the citation count of a paper. Well-cited papers will become more well cited. If the total number of citations per paper grew to accommodate the increasing number of papers as a field grows, then this inequality might not be inevitable. But reference lists do not get ever longer and, as a result, the frequency of paper citation counts tends to follow a rank-size pattern*". In the case of top cited papers in this study, they belong to highly visible academics of the large, established and dominant disciplinary field, whose articles in general and the Zipf ones in particular, generate hundreds to thousands of citations (2133 for Gabaix's 1999 "Zipf's law for cities, an explanation"). Finally Nitsch (2005)'s meta analysis is frequently cited (21% of all sample articles but 34% of those published in or after 2005). Many externally cited references do not appear on this graph for they receive less than 5 mentions from the 66 corpus bibliographies[1].

3.3. Translating journals into disciplines/disciplinary fields

In order to study the disciplinary dynamics of such works on Zipf's law for cities, we finally assigned a field to each of the 707 journals and publishing institutions from which internal and external references of this meta meta analysis were taken. We chose to identify 5 fields: Economics (ECO), Geography (GEO), Regional Science and planning (REG), Statistics (STAT) and Physics (PHY). Although identification of journals in the last two fields was rather straightforward, the lines between Economics, Geography and Regional Science were quite blurry. However, it seemed interesting to separate the three for two reasons. Firstly, economics and geography are recognised disciplines whose practitioners do not frequently publish in each other's journals, whereas Regional Science sits precisely at the intersection between economics, geography and planning. In regional sciences/studies conferences and journals, it is not usually to find legacies and references to both fields. We thus wanted to be able to assess this features in the sub-field of city size distribution studies. Secondly, the separation acknowledges the fact that publication strategies vary greatly between the journals of these fields, in terms of exposure, sphere of impact, formal and theoretical requirements, etc. even when articles deal with the same object (the city size distribution of cities).

The key use to affect journals between the three fields were the following:

- "ECO" for general economics journals (such as the Quarterly Journal of Economics) as well as journals with "economics" in their names (Journal of Urban Economics for instance)
- "REG" when the subject is "urban affairs", "urban studies", or has "urban and regional" in the name.
- "GEO" for general geography journals (for example, the Annals of the Association American Geographers) as well as journals with names referring to the processes of urban development and urbanisation.

This approach contains some ad-hoc character. We have tried to alleviate it by providing access to the lookup table to engage conversations with potential readers in disagreement. We are aware of existing journal classifications but find that they do not reflect entirely the stakes of this sub-field (nor do they provide guidance for the classification of books, reports and dissertation). An assessment of the most frequent journals with the Scimago classification shows for instance that the ad-hoc fields we attributed to journals always match at least one of the Scopus subject areas

1 Some of them are indeed quite specific, for instance those from the aerosol literature cited in to Eeckout (2004): **Haaf, Amin and Jaenicke, Rainer.** "Results of Improved Size Distribution Measurement in the Aitken Range of Atmospheric Aerosols." *Journal of Aerosol Science*, 1980, *11*(3), pp. 321–30. & **Hinds, William C.** *Aerosol technology*. New York: Wiley, 1982.

proposed by Scimago[2], considering that "Environmental Science (miscellaneous)" corresponds to Regional Science and planning (table S1 in supplement). The advantage of our system is that it provides a single category for each source, whereas Scimago has a varying number of entries for different journals, and no entry for French journals like "Région et Développement" which is externally cited 6 times in our corpus, or for dissertations and World bank databases.

Table 1. Distribution of references by discipline of the journal they were published in.

Disciplinary field	ECO	GEO	STAT	REG	PHY	OTHER
Corpus	23	13	0	26	4	0
External references	341	233	126	125	49	281

After applying this ad-hoc translation to all external reference, we can see a stark difference between the distribution of corpus stydies by disciplinary field and that of their bibliography (table 1). Indeed, while most corpus are published in regional science and economics journals, but their framework of reference comes primarily from economics and geography, or at least articles published in economics and geography journals. Secondarily, corpus articles draw from the statistics (for estimation methods and tools) and regional science literatures. Thirdly they cite articles published in physical science journals. The large number of "OTHER" references indicates the diversity of Zipf-related work bibliographies, which frequently reference other disciplines (political science, architecture, etc.), other formats (reports, dissertation, etc.) and languages.

3.4. From individual studies to reference networks

In order to assess whether the difference in meta properties of articles dedicated to the empirical estimation of Zipf's law play a role in the variation of results they report, we constructured seven networks of similarity. The seven networks have each 66 vertices corresponding to every corpus articles. They differ in the distribution of edges connecting vertices. The first three networks ("wording", "citation" and "disciplinary") were built to test our three hypotheses. The next three network ("country", "decades" and "city definition") were built to controle for the similarity of the objects actually studied by corpus articles. The last network ("alpha") is the one to enventually "explain": it is the network of corpus articles drawn by the similarity of the distribution of Zipf estimates they report.

Similarity for all networks was measured pairwise, by the cosine similarity of two corpus articles' vectors. A visual representation of each network is performed, using the 'igraph' R package (Csardi & Nepusz, 2006). For better visibility, we apply a cutoff to the weight of edges represented and exclude non-connected vertices and colour vertices using Louvain community clusters (except in figure 6). These representations provide clues for interpretation. However, the modelling analysis is run on the complete network. The entire analysis is made available on Github[3].

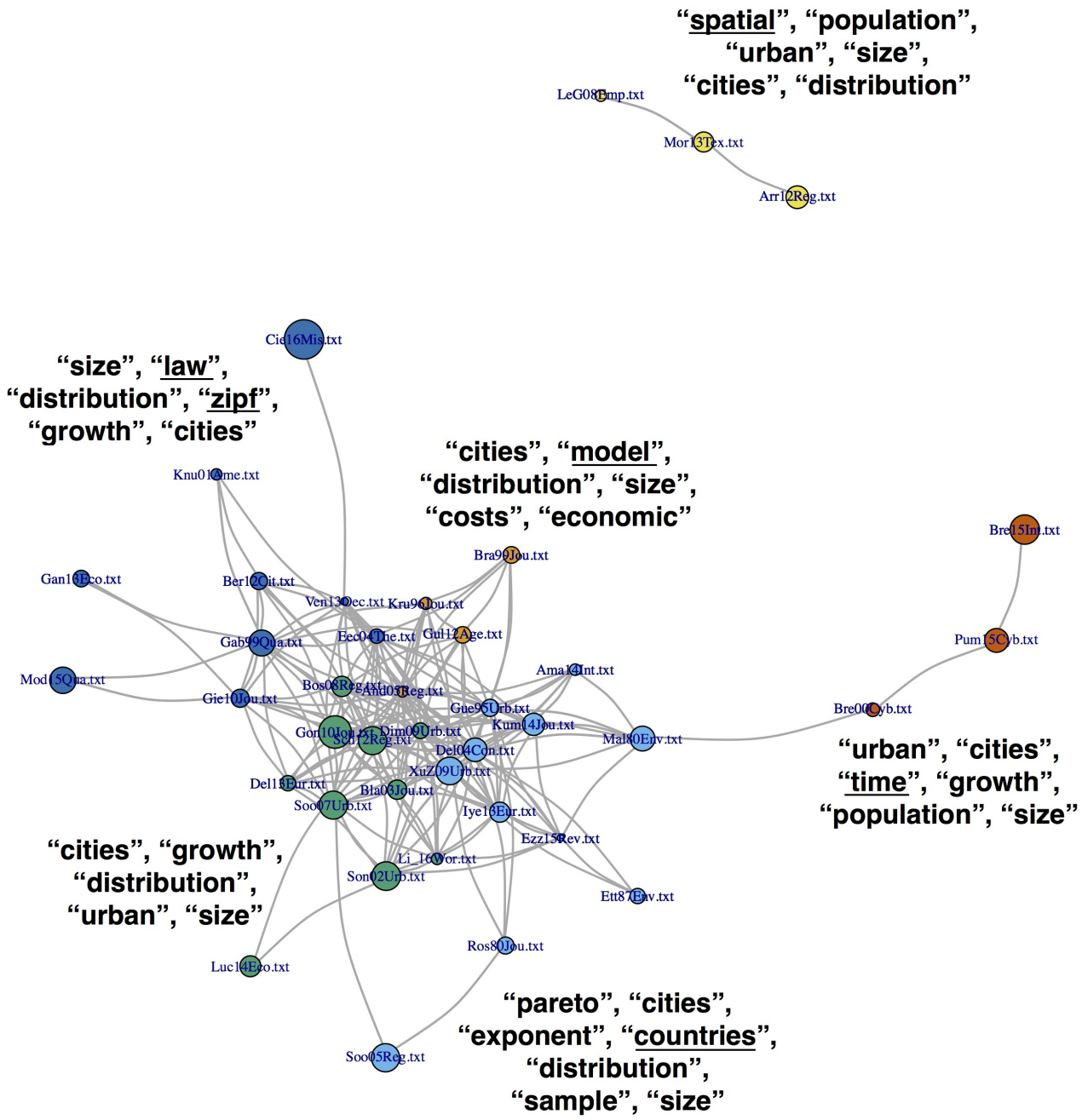
3.4.1. The "wording" network.

The "wording" network represents the similarity between corpus articles based on the frequency of words they have used to write their paper and present empirical estimations of Zipf's exponents. Using the 66 full texts collected, we computed the frequency distribution of 10,791 non-stop words in each corpus articles. The vectors used as inputs for the "wording" cosine similarity are thus composed of 10,791 values comprised between 0 and 1. A visualisation of a subset of the network created is visible on figure 4, along with some of the most frequent terms used by corpus articles of the different communities. The variation is vertices sizes represents their total number of terms.

2 https://www.scimagojr.com/ (accessed on 20/08/2020)

3 https://github.com/ClementineCttn/MetaZipf

Figure 4. Similarity network of corpus articles by frequency of words used (cut-off at 0.65).

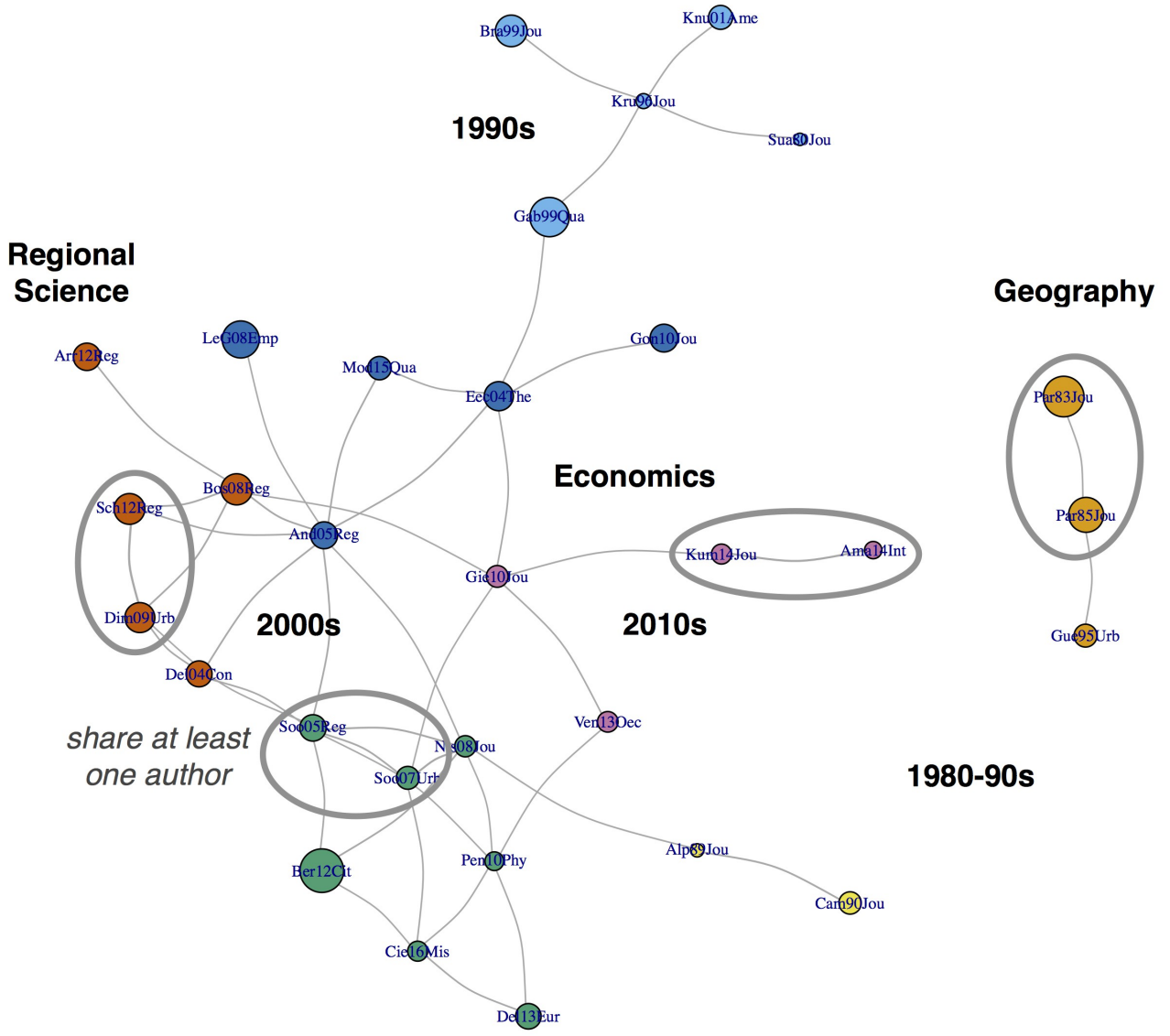


The figure shows a network with strong connectivity. Indeed, most articles use, at the very least, the words "cities", "size" and "distribution" very frequently. However, a disconnected community of three articles (Le Gallo & Chasco 2008; Moro & Santos, 2013 et Arribas-Bel et al., 2012, in yellow) shows a more important use of the word "spatial". These works even have "spatial" in their title. Their aim is not to verify Zipf's law but to present and analyse a national urban system, respectively Spain, Brasil and Australia. Another cluster (in red) shows similar of wording with the particularity of using "time" very often. Originating from a unique research group in France, Bretagnolle et al., (2000, 2015) and Pumain et al. (2015) indeed present long-term evolutions of systems of cities, reporting on their growth and structure of several decades. The light blue cluster gathers comparative studies who therefore make a more thourough use of the term "countries". Two other clusters represent articles less devoted to empirical analysis and more to the testing of "Zipf"'s "law" (dark blue) or "model"-ing its generation (organge). This network thus represent the way Zipf's law is approached by authors and the finality of the argument.

3.4.2. The “citation” network.

The “citation” network represents the similarity between corpus articles based on the external references they cite in the course of the text. It could be argued that two papers citing the exact same corpus of references (high similarity) could share the same aim, such as “proving” or “disproving Zipf's law”, and therefore report similar estimate values. The similarity was measured for each pair of the 66 vectors of 1155 external references, coded 1 if the reference was cited and 0 otherwise. A visualisation of a subset of the network created is visible on figure 5, with the total number of external citations being represented by the size of vertices.

Figure 5. Similarity network of corpus articles by external articles cited (cut-off at 0.25).

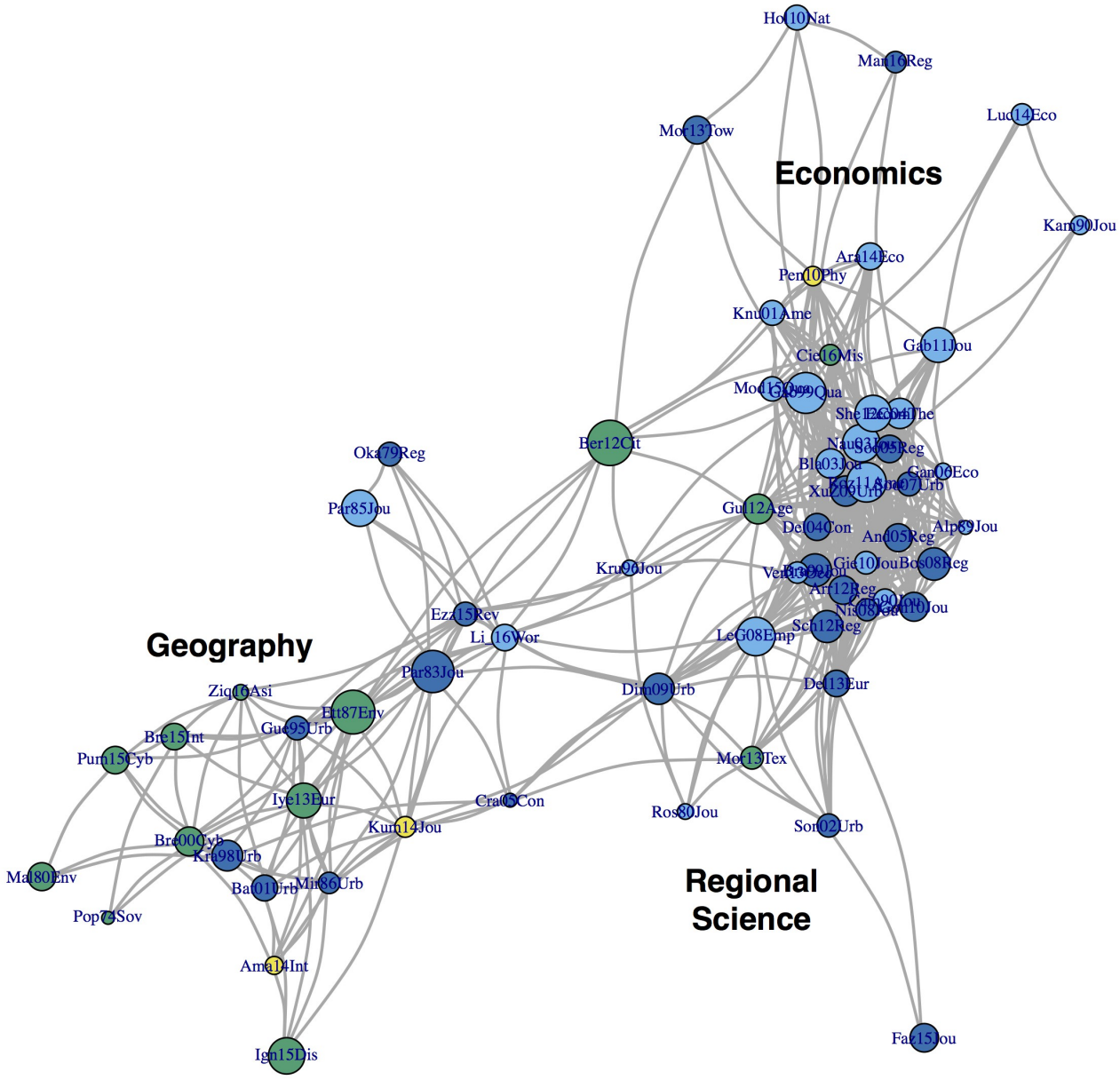


This network is less connected than others, suggesting that the reference framework of every author depends on a diversity of causes besides a common object of study. Indeed, the subnetwork shown in figure 5 is organised along periods, co-authorship and disciplinary lines. The similarity of citations is, quite trivially, harder for articles of different periods because of the unavailability of later references for earlier articles. Therefore, we see clusters of corpus articles along publication dates (orange and yellow clusters in the 1980s and early 1990s, light blue cluster in the late 1990s, red and green clusters in the 2000s and 2010s, pink cluster in the 2010s). The co-authored articles (Soo, 2005 and 2007 or Dimou & Schaffar, 2009 and Schaffar & Dimou, 2012 for instance) reveal from the inertia of authors' reference framework over time and the individualised take on article subjects. Finally, articles published in economics journals seem to share more bibliography than

they do with articles published in geography, with regional science closer to economics in that respect.

3.4.3. The “discipline” network.
The “discipline” network represents this aspect more broadly, as it represents the similarity between corpus articles based on the discipline of the journal their external references were published in. The cosine similarity was measured on vectors of 6 items (the number of external refences from each discipline). For this representation, we did not use community clusters for colouring nodes but instead the discipline of the journal where corpus articles were published.

Figure 6. Similarity network of corpus articles by external disciplines cited (cut-off at 0.9).



** Here the colour of the nodes shows the discipline of the node's article discipline rather than the membership to a Louvain community cluster. Yellow: physics. Green: geography. Light blue: economics. Dark blue: regional science and planning.*

The figure shows that some entanglement of disciplinary references, with regional science corpus articles citing a similar pool of disciplines as geography corpus articles. However, this might be an artifact of publication strategies, since the articles in question (such as Parr & Jones, 1983, Guérin-

Pace, 1995 or Batty, 2001) are authored by people recognised as geographers. On the other hand, articles in economics and in regional science also share similar disciplinary references. The divergence of disciplinary framework appears main between geography and economics articles, although some articles (Krugman, 1996; Berry and Okulicz-Kozalyn, 2012 or Dimou & Schaffar, 2009) work as bridges, citing from a more varied pool of disciplinary references.

3.4.4. The “country” network.

The “country” network represents the similarity between corpus articles based on the countries on which they perform empirical estimations of Zipf's exponents (figure S2). High similarity characterise articles dedicated to the same area (USA, China, South Africa) and articles dedicated to comparative studies (like the most extensive of that kind: Soo, 2005; Rosen & Resnick, 1980). The two densest clusters are composed by studies reporting Zipfùs estimates exclusively for American (in orange) and Chinese (in blue) cities.

3.4.5. The “decades” network.

The “country” network represents the similarity between corpus articles based on the decades on which they perform empirical estimations of Zipf's exponents (figure S3). High similarity characterise articles dedicated to the same period. The densest clusters are composed by corpus articles reporting Zipf's estimates exclusively for a single decade (such as Cameron, 1990 or Krugman, 1996)

3.4.6. The “city definition” network.

The “city definition” network represents the similarity between corpus articles based on the city definition used to collection city populations (municipality, agglomeration or metropolitan areas mostly) on which empirical estimations of Zipf's exponents are performed (figure S4). This network is polarised by the use of one or more city definition in the corpus article. The larger cluster (in orange) unfortunately reflects the fact that most city size distribution are analysed for improper urban delineations (the 'city proper' or municipal boundaries), as their shape and principles vary greatly across countries but tend to stay fixed over time whereas cities expand spatially and functionally.

3.4.7. The “alpha” networks.

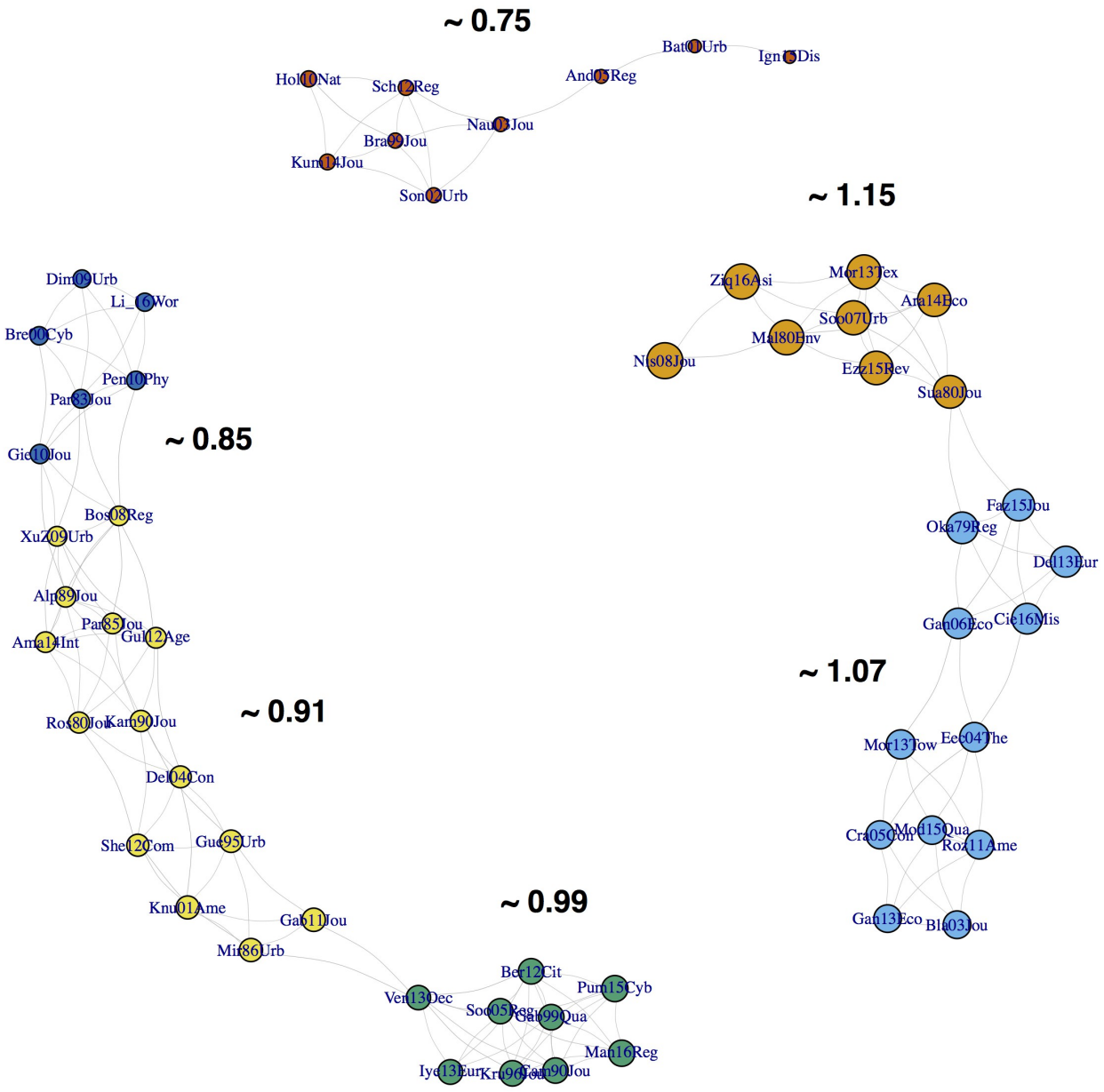
The “alpha” networks represents the similarity between corpus articles based on the distribution of empirical estimations of Zipf's exponents (alpha expressed under the Lotka form, or 1/alpha expressed in the Pareto form) they report. We choose to model two aspects of this distribution: the average value of alpha reported on the one hand, and its standard deviation on the other hand. Additionally, we use the number of estimates reported as an extra control.

To construct the “mean alpha” network, we compute the average value of alpha estimates $\bar{a}_i$ per study i and the average value of alpha estimates $\bar{a}$ for the entire sample (1962 estimates). We then compute a distance $d\bar{a}_{ij}$ between studies as follows:

$$d\bar{a}_{ij} = | \bar{a}_i - \bar{a}_j | / \bar{a} \quad , \text{with } i \neq j$$

The smaller this distance the more studies i and j report Zipf estimates close in value. To transform this distance into a similarity, we simply multiply $d\bar{a}_{ij}$ by -1. The network emerging from this similarity is therefore organised around groups of studies based on the average values of alpha estimates they report (figure 7). At the low end of the sprectrum, studies like Holmes & Lee (2010) or Kumar & Subbarayan (2014) report very low values of estimates (0.75 on average for the group in red), which indicates city sizes more evenly distributed than predicted by Zipf. At the other end of the spectrum, studies like Ziqin (2016) or Nishiyama et al. (2008) report high values of estimates (1.15 on average for the group in orange), which reflects highly uneven city size distributions.

Figure 7. Similarity network of corpus articles by average value of estimates reported (cut-off at -0.025).



* the size of nodes reflects the average value of estimates reported in the article and the numbers in black correspond to the average value reported for the community.

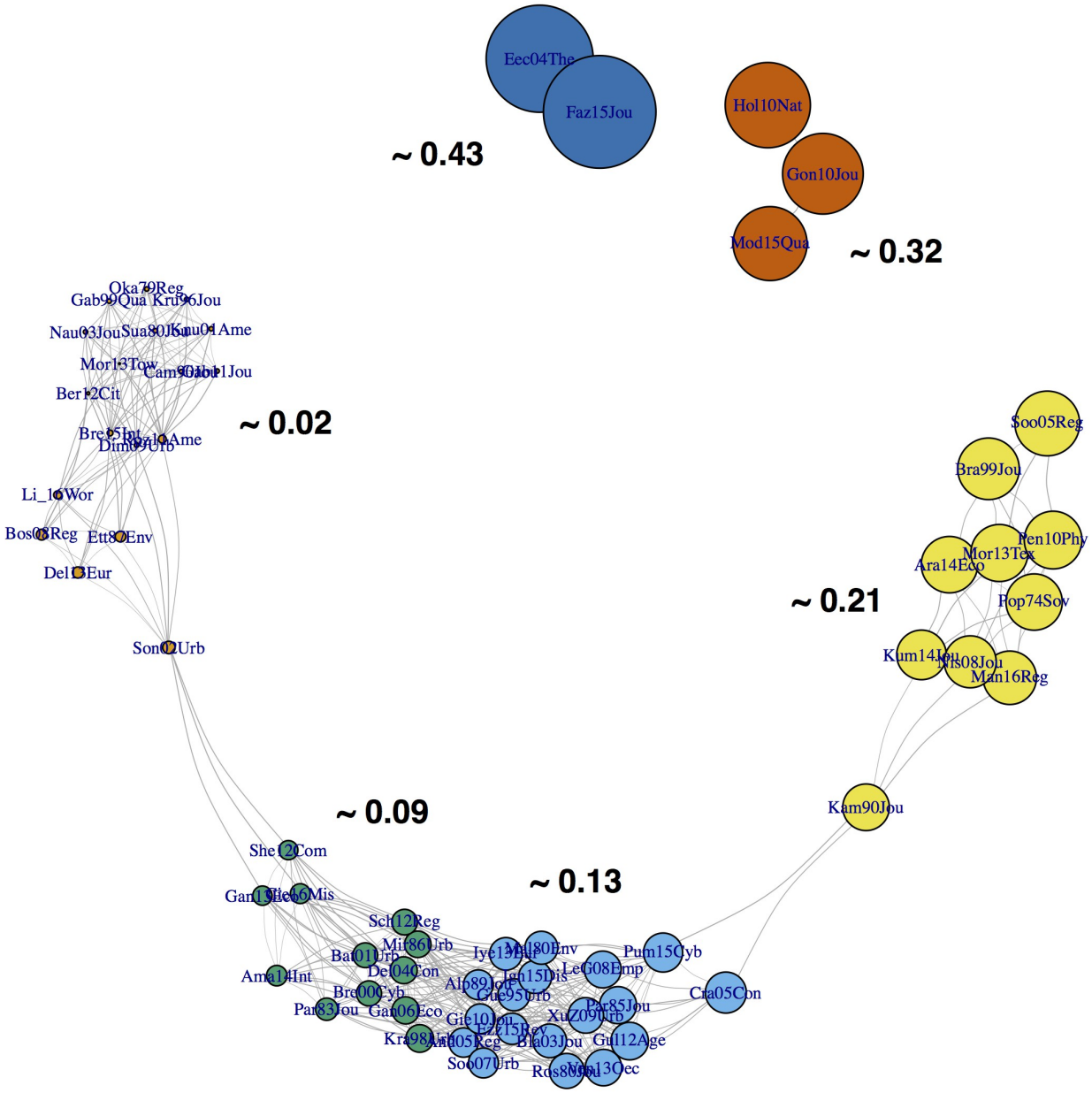
To construct the "standard deviation" network, we compute the standard deviation σ^2_i of alpha estimates per study i and the standard deviation of alpha estimates σ^2_a for the entire sample. We then compute a distance $D\sigma^2_{ij}$ between studies as follows:

$$D\sigma^2_{ij} = | \sigma^2_i - \sigma^2_j | / \sigma^2_a \quad , \text{with } i \neq j$$

The smaller this distance the more studies i and j report a similar dispersion of Zipf estimates. To transform this distance into a similarity, we simply multiply $D\sigma^2_{ij}$ by -1. The network emerging from this similarity is therefore organised around groups of studies based on the average dispersion of alpha estimates they report (figure 8). At the low end of the sprectrum, studies like Okabe (1979) or Gabaix (1999) report estimates very close to one another (0.02 standard deviation on average for the group in orange), frequently because such studies only report 1 or 2 estimates. At the other end of

the spectrum, studies like Eeckout (2004) or Fazio & Modica (2015) report very dispersed distributions of estimates (0.43 standard deviation on average for the group in dark blue). In these two examples, such dispersion is produced by estimations all made for the USA in 2000 and 2010, but with large variations of truncation points (i.e. the minimum population size to include cities in the sample), from 135 residents to 29,000, which changes the number of places included in the regression from about 156,000 to only 35. AS noted in Cottineau (2017), the truncation point is one of the most important technical specifications with respect to the variation of Zipf's estimates in the literature.

Figure 8. Similarity network of corpus articles by standard deviation of estimates reported (cut-off at -0.1).



* the size of nodes reflects the standard deviation of estimates reported in the article and the numbers in black correspond to the average standard deviation for the community.

Finally, we constructed a “n alpha” network to control for the number of estimates reported (especially when modelling their dispersion). We computed the number alpha estimates n_i per study and the average value of alpha estimates $\bar{n}$ in the entire sample. We then computed a distance dn_{ij} between studies as follows:

$$Dn_{ij} = | n_i - n_j | / \bar{n} \quad , \text{with } i \neq j$$

The smaller this distance the more studies i and j report a similar number of Zipf estimates. To transform this distance into a similarity, we simply multiply $d\bar{a}_{ij}$ by -1. The network emerging from this similarity is shown in suppementary figure S3.

3.5. Modelling dyad similarities

We run two series of models, one aimed at "explaining" the similarity in mean alpha values reported between corpus articles, and one aimed at explaining their similar dispersion. "Explaning" variables for each series of models are similarity measures of the "wording", "citation", "disciplinary", "country", "decades", "city definition" and "n alpha" networks. All variables were centered and scaled prior to modelling. We analyse the coefficients b of the models as well as their residuals e_{ij}.

$$MeanAlpha_{ij} = b_1 Wording_{ij} + b_2 Citation_{ij} + b_3 Discipline_{ij} + b_4 nAlpha_{ij} + b_4 Country_{ij} + b_5 Decade_{ij} + b_6 CityDef_{ij} + e_{ij}$$

$$sdAlpha_{ij} = b_1 Wording_{ij} + b_2 Citation_{ij} + b_3 Discipline_{ij} + b_4 nAlpha_{ij} + b_4 Country_{ij} + b_5 Decade_{ij} + b_6 CityDef_{ij} + e_{ij}$$

with i≠j, i and j being articles from the corpus.

We also look at interactions between control variables to identify studies which have studied similar national systems in the same decade and under the same definition of cities. These studies should report the most similar rank-size estimations.

4. Results

The results of such modelling is reporting in tables 2 and 3. Regarding the similarity in mean alpha reported in the corpus (table 2), we find a confirmation of two of our three initial hypotheses. Although the R^2 are low, the similarity in mean alpha varies positively and significantly with both the similarity in wording and the similarity in citations (models 1, 2 and 5). This means that articles written with a similar set of words and references tend to report similar values of Zipf estimates on average. This interesting feature persists (model 5) even when we account for the similarity in countries, decades and city definitions which the pair of corpus articles studies. As the wording network showed, this could result from a different setup from which the estimation originates. In some articles, the goal is to validate a "law" and the adequacy of one case to the "model". It is thus more probable that such studies report estimates centered around -1, as in the strict version of Zipf's law. On the other hand, articles citing the same pool of references can exhibit a similar interest in validating or challenging the law. The evidence for the similarity in disciplinary references is more mixed, since the significant effect of the simpe model 3 disappears when other variables and controls are accounted for. In terms of controls, we find that articles reporting a similar number of estimates then to differ in mean alpha. This can be the effect of sensitivity studies which explore the effect of threshold values or other specification criteria: they generally report a high number of estimates but their dispersion is such that the average value varies a lot. As expected, studies which analyse the same set of countries tend to report similar values of estimates on average, however the opposite is true for time periods. The effect of similar city definitions chosen to analyse siwe distributions was not found significant by itself but appeared positive in conjunction with a similarity in the set of countries and with a similarity in the set of decades studied, as expected.

Regarding the similarity in dispersion (table 3), we find that only one of our main hypotheses is verified: the more articles are written with similar words, the more similar they are in terms of standard deviation of alphas reported (models 1 and 6). Again, some articles are similar in their attempts at verifying the "law": they are written with mathematical language and tend to report few estimates close in value. Other articles have the goal of exploring the national variation of city size distributions or their sensitivity to technical specifications: they use words like "countries", "spatial" and "comparison" and tend to report a very dispersed set of results.

Table 2. OLS regression of the similarity in average value of alpha reported in the corpus:

	Dependent variable:				
	similarity in meanAlpha				
Similarity in ...	(1)	(2)	(3)	(4)	(5)
wording	0.048**				0.050**
	(0.022)				(0.023)
citation		0.062***			0.069***
		(0.022)			(0.024)
discipline			0.043*		0.015
			(0.022)		(0.024)
nAlpha				-0.053**	-0.046**
				(0.022)	(0.022)
country				0.063***	0.063***
				(0.023)	(0.023)
decade				-0.100***	-0.123***
				(0.022)	(0.023)
cityDef				-0.001	-0.015
				(0.022)	(0.023)
country:decade				-0.014	-0.013
				(0.021)	(0.021)
country:cityDef				0.049**	0.053**
				(0.022)	(0.022)
decade:cityDef				0.037*	0.038*
				(0.022)	(0.022)
country:decade:cityDef				-0.017	-0.014
				(0.021)	(0.021)
Constant	0.015	0.015	0.015	0.013	0.012
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Observations	2,016	2,016	2,016	2,016	2,016
R^2	0.002	0.004	0.002	0.021	0.030
Note:					*p<0.1; **p<0.05; ***p<0.01

The distribution of positive residuals (figure S4) shows similarity between pairs higher than estimated by the model. No obivous pattern seem to govern the association between such pairs, whereas luck might play a role. However, negative residuals are driven by three studies whose average estimate value differ from that of all others: Luckstead & Devadoss (2014), Le Gallo & Chasco (2008) and Popov (1974). They report an average value alpha respectively of 1.91, 1.73 and 1.45. Those are very far away from the expected linear exponent of Zipf's law, which might suggest to consider outlyers for a subsequent meta analysis of the empirical literature.

Table 3. OLS regression of the similarity in standard deviation of alpha reported:

	Dependent variable:					
	Similarity in sdAlpha					
Similarity in ...	(1)	(2)	(3)	(4)	(5)	(6)
wording	0.112***					0.116***
	(0.022)					(0.023)
citation		-0.013				-0.006
		(0.022)				(0.024)
discipline			-0.004			-0.009
			(0.022)			(0.024)
nAlpha				0.134***		0.133***
				(0.022)		(0.022)
country					-0.096***	-0.078***
					(0.023)	(0.022)
decade					-0.077***	-0.076***
					(0.022)	(0.023)
cityDef					0.088***	0.083***
					(0.022)	(0.022)
country:decade					0.041*	0.042**
					(0.021)	(0.021)
country:cityDef					0.050**	0.051**
					(0.022)	(0.022)
decade:cityDef					0.050**	0.051**
					(0.022)	(0.022)
country:decade:cityDef					-0.023	-0.016
					(0.021)	(0.021)
Constant	-0.009	-0.009	-0.009	-0.008	-0.014	-0.013
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Observations	2,016	2,016	2,016	2,016	2,016	2,016
R^2	0.012	0.0002	0.00002	0.018	0.027	0.055

We do not find any significant evidence of covariation between the similarity in bibliography and displines cited and the similarity in alpha dispersion. However, the number of estimates is shown to positively influence the similarity in dispersion, since more estimated tends to increase the dispersion on average. Studies which use similar city definitions tend to report similar dispersion. Finally, although the similarity in countries and decades studied is negatively associated with a similarity in dispersion per se, they are positively associated when in interaction with one another and with city definition (model 6). The distribution of residuals (figure S5) exhibits the same properties as that of the previous model: elective similarity between more or less isolated pairs of studies and polarised dissimilarity with a couple of articles, including Luckstead & Devadoss (2014).

5. Discussion & Conclusion

In this article, we have looked at the empirical literature on Zipfs law for cities from a network perspective. As a complement to previous meta-analyses, the present approach has shed light on the scientific text and context mobilized to report on city size distributions. As in Raimbault et al., 2019, it has used textual analysis and citation networks to reflect various proximities between articles of the corpus. The analysis of each network had produced insight in the wording, reference framework and disciplinary heritage demonstrated by the empirical literature on Zipfs law for cities. Their use as explaining variables of a model of the similarity in the distribution of estimates reported has shown that wording is important in both cases, whereas similar citation patterns mostly impact the average value of Zipf's estimate reported.

The contribution of this paper to meta-analyses has been two-fold. Firstly, using the citation networks of studies included in a meta-analysis has allowed us to identified gaps in the corpus and potentially overlooked articles. These have appeared when looking at the most cited external references. In our case, the article of Eaton & Eckstein (1997) for example is one of the most externally cited reference to report empirical estimations of Zipf's law. It was initially rejected from the corpus (Cottineau, 2017) because the estimation included instruments. The present analysis suggests that relaxing this criterion could allow its inclusion as a major reference in the field. Symmetrically, the analysis of model residuals has shown that some very atypical studies drive a large share of the difference in mean values and dispersion used in the meta-analysis, suggesting that removing them as outlyers could provide clearer results. Secondly, the data and code of the present study has been made open on github, including an R notebook with all visualisations, in order to be reused by the community.

Although this article does not close the debate on city size distribution, it has tried to reveal a newer aspect of a literature in rapid development: the fact that it mixes studies of very different aims and methods, potentially characterised by reporting biases. What seems quite obvious from the corpus is also the fact that Zipf's law estimation is a large field where many authors contribute at one point of their scientific career in urban studies, economics or physics, but mostly not an object of research per se. A further point of inquiry in the reflexive meta-analysis could thus be to trace various authors' contribution to the empirical Zipf literature as part of their scientific topic trajectory (Zeng et al., 2019). However, it is not obvious at this point to which extend it would help provide guidelines for rigorous analysis of city size distribution.

6. References

Arribas-Bel, D., Gracia, F. S., & Ximénez-de-Embún, D. (2012). Kangaroos, cities and space: a first approach to the australian urban system. Region et Developpement, 36, 165–187.

Auerbach F. (1913) Das Gesetz der Bevölkerungskonzentration. Petermanns Geographische

Mitteilungen;59:74–76.
Berry, B. J., & Okulicz-Kozaryn, A. (2012). The city size distribution debate: Resolution for US urban regions and megalopolitan areas. *Cities*, *29*, S17-S23.
Bretagnolle, A., Mathian, H., Pumain, D., & Rozenblat, C. (2000). Long-term dynamics of European towns and cities: towards a spatial model of urban growth. Cybergeo: European Journal of Geography. http://doi.org/10.4000/cybergeo.566
Bretagnolle, A., Delisle, F., Mathian, H., & Vatin, G. (2015). Urbanization of the United States over two centuries: an approach based on a long-term database (1790–2010). International Journal of Geographical Information Science, 1–18.
Cameron, T. A. (1990). One-stage structural models to explain city size. Journal of Urban Economics, 27(3), 294–307.
Christaller, W. (1933). Die zentralen Orte in Süddeutschland, *Jena: Gustav Fischer*.
Cottineau, C. (2017). MetaZipf. A dynamic meta-analysis of city size distributions. *PloS one*, *12*(8), e0183919.
Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, complex systems*, *1695*(5), 1-9.
Dimou, M., & Schaffar, A. (2009). Urban hierarchies and city growth in the Balkans. Urban Studies. http://usj.sagepub.com/content/early/2009/09/04/0042098009344993.short
Eaton, J., & Eckstein, Z. (1997). Cities and growth: Theory and evidence from France and Japan. *Regional science and urban Economics*, *27*(4-5), 443-474.
Eeckhout, J. (2004). Gibrat's Law for (All) Cities. The American Economic Review, 94(5), 1429–1451.
Fazio, G., & Modica, M. (2015). Pareto or Log-Normal? Best Fit and Truncation in the Distribution of All Cities*. Journal of Regional Science, 55(5), 736–756. http://doi.org/10.1111/jors.12205
Feinerer I., Hornik K., Artifex Software Inc., 2019, "tm: Text Mining Package", R package, version 0.7-7, https://CRAN.R-project.org/package=tm
Gabaix, X. (1999). Zipf's law for cities: anexplanation. Quarterly Journal of Economics, 739–767
Gabaix, X., & Ioannides, Y. M. (2004). The evolution of city size distributions. In *Handbook of regional and urban economics*, Vol. 4, pp. 2341-2378
Gibrat, R. (1931). Les inégalités économiques. Paris, Libraire du Recueil Sierey.
Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 1163-1174.
Holmes, T. J., & Lee, S. (2010). Cities as Six-by-Six-Mile Squares: Zipf's Law? NBER, 105–131
Ignazzi, C.A., 2015, Coevolution in the Brazilian system of cities, Dissertation, Université Paris 1, Panthéon-Sorbonne, Paris, France
Kumar, G., & Subbarayan, A. (2014). The temporal dynamics of regional city size distribution: Andhra Pradesh (1951-2001). *Journal of Mathematics and Statistics*, 10(2), 221
Le Gallo, J., & Chasco, C. (2008). Spatial analysis of urban growth in Spain, 1900–2001. Empirical Economics, 34(1), 59–80.
Lösch, A. (1940). Die räumliche Ordnung der Wirtschaft: eine Untersuchung über Standort. *Gustav Fisher, Jena*.
Luckstead, J., & Devadoss, S. (2014). A comparison of city size distributions for China and India from 1950 to 2010. Economics Letters, 124(2), 290–295. http://doi.org/10.1016/j.econlet.2014.06.002
Moro, S., & Santos, R. (2013). The characteristics and evolution of the Brazilian spatial urban system: empirical evidences for the long-run, 1970-2010(Textos para Discussão Cedeplar-UFMG No. 474). Cedeplar, Universidade Federal de Minas Gerais. Retrieved from http://econpapers.repec.org/paper/cdptexdis/td474.htm
Nishiyama, Y., Osada, S., & Sato, Y. (2008). OLS estimation and the t test revisited in rank-size rule regression. *Journal of Regional Science*, 48(4), 691–716.
Nitsch, V. (2005). Zipf zipped. *Journal of Urban Economics*, *57*(1), 86-100.

Okabe, A. (1979). An expected rank-size rule: A theoretical relationship between the rank size rule and city size distributions. Regional Science and Urban Economics, 9(1), 21–40.
Pareto, V. (1897). Cours d'économie politique, Vol. 2. *Lausanne*, Rouge.
Parr, J. B. (1985). A note on the size distribution of cities over time. *Journal of Urban Economics*, *18*(2), 199-212.
Popov, V. R. (1974). Investigation of the System of Urban Places of Crimea Oblast. Soviet Geography, 15(1), 19–23.
Pumain, D., Swerts, E., Cottineau, C., Vacchiani-Marcuzzo, C., Ignazzi, A., Bretagnolle, A., ... Baffi, S. (2015). Multilevel comparison of large urban systems. Cybergeo: European Journal of Geography. http://doi.org/10.4000/cybergeo.26730
Raimbault, J., Chasset, P. O., Cottineau, C., Commenges, H., Pumain, D., Kosmopoulos, C., & Banos, A. (2019). Empowering open science with reflexive and spatialised indicators. *Environment and Planning B: Urban Analytics and City Science*, 2399808319870816.
Rosen, K. T., & Resnick, M. (1980). The size distribution of cities: an examination of the Pareto law and primacy. Journal of Urban Economics, 8(2), 165–186.
Schaffar, A., & Nassori, D. (2016). La croissance urbaine marocaine: convergence vs concentration. Revue économique, 67(2), 207–226.
Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425-440.
Singer, H. W. (1936). The" courbe des populations." A parallel to Pareto's Law. *The Economic Journal*, *46*(182), 254-263.
Soo, K. T. (2005). Zipf's Law for cities: a cross-country investigation. Regional Science and Urban Economics, 35(3), 239–263.
Soo, K. T. (2007). Zipf's Law and urban growth in Malaysia. Urban Studies, 44(1), 1–14.
Suarez-Villa, L. (1980). Rank size distribution, city size hierarchies and the Beckmann model: some empirical results. Journal of Regional Science, 20(1), 91–95
Webster C. (2006) "Editorial. Ranking planning journals", *Environment and planning B*, volume 33, pages 485-490, https://journals.sagepub.com/doi/pdf/10.1068/b3304ed
Woolgar, S., & Latour, B. (1986). *Laboratory life: the construction of scientific facts*. Princeton University Press.
Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature communications*, *10*(1), 1-11.
Zipf, G. K. (1941). National Unity and Disunity, Blomington, Principia Press.
Zipf, G. K. (1949). *Human behavior and the principle of least effort*. New York: Hafner, 573.
Ziqin, W. (2016). Zipf Law Analysis of Urban Scale in China. *Asian Journal of Social Science Studies*, 1(1), 53

Supplementary Material

Table S1. Correspondance between ad-hoc fields and scimago classification of the most externally cited journals (at least five citations from the corpus).

Journal	Cites	Group	Scopus Subject Area & category 1	Scopus Subject Area & category 2
Journal Regional Science	30	REG	Environmental Science	Development
Urban Studies	26	REG	Environmental Science	Urban Studies
Journal Urban Economics	25	ECO	Economics and Econometrics	Urban Studies
American Economic Review	16	ECO	Economics and Econometrics	x
Economic Development Cultural Change	15	ECO	Economics and Econometrics	Development
Econometrica	14	ECO	Economics and Econometrics	x
Environment Planning A	14	GEO	Geography, Planning and Development	Environmental Science
Quarterly Journal Economics	14	ECO	Economics and Econometrics	x
Regional Science Urban Economics	14	REG	Urban Studies	Economics and Econometrics
International Regional Science Review	11	REG	Environmental Science (miscellaneous)	Social Sciences
Physica A	11	PHY	Condensed Matter Physics	Statistics and Probability
Papers in Regional Science	15	REG	Environmental Science	Geography, Planning and Development
Annals AAG	9	GEO	Geography, Planning and Development	Earth-Surface Processes
Geographical Analysis	9	GEO	Geography, Planning and Development	Earth-Surface Processes
Journal Political Economy	9	ECO	Economics and Econometrics	x
Journal Econometrics	8	ECO	Economics and Econometrics	Applied Mathematics
Journal Economic Geography	8	ECO	Economics and Econometrics	Geography, Planning and Development
NBER	8	ECO	x	x
Annals Regional Science	6	REG	Environmental Science	Social Sciences
CEPR	6	ECO	x	x
Dissertation	6	OTHER	x	x
Geographical Review	6	GEO	Geography, Planning and Development	Earth-Surface Processes
Journal American Statistical Association	6	STAT	Statistics, Probability and uncertainty	Statistics and Probability
Region Development	6	REG	x	x
Journal Royal Statistical Society	5	STAT	Statistics and Probability	Economics and Econometrics
Physical Review E	5	PHY	Condensed Matter Physics	Statistical and Nonlinear Physics
Professional Geographer	5	GEO	Geography, Planning and Development	Earth-Surface Processes
Regional Studies	5	REG	Environmental Science	Social Sciences
Review Economic Studies	5	ECO	Economics and Econometrics	x
World Development	5	OTHER	Development	Sociology and Political Science
WorldBank	5	STAT	x	x

Figure S1. Distribution and density of corpus articles over the years, according to their citation of Zipf's works or not.

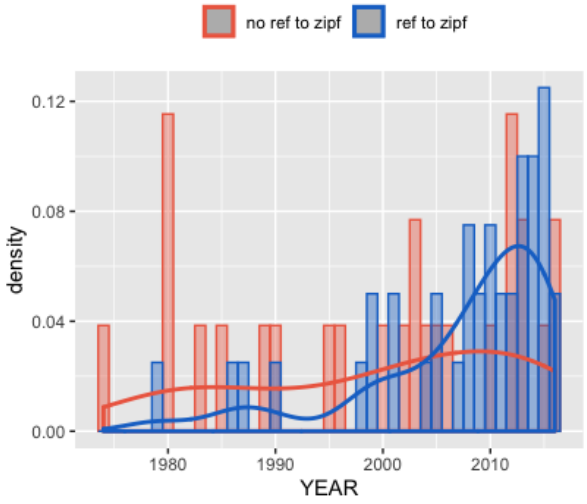


Figure S2. Similarity network of corpus articles by the common countries they reported alpha on (cut-off 0.25). The size of vertices represent the number of countries they report estimates for.

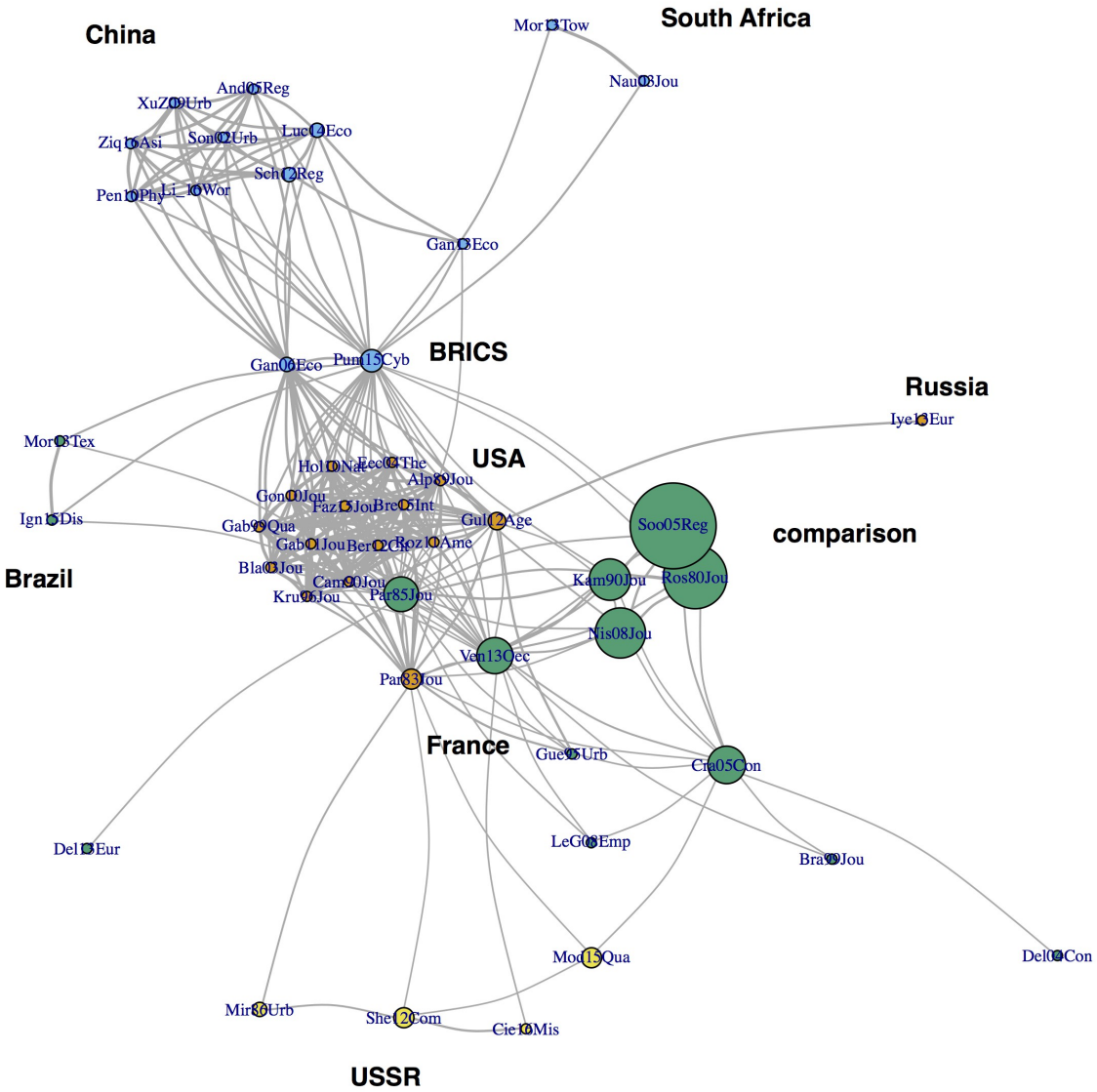


Figure S3. Similarity network of corpus articles by the common decades they reported alpha on (cut-off at 0.65).

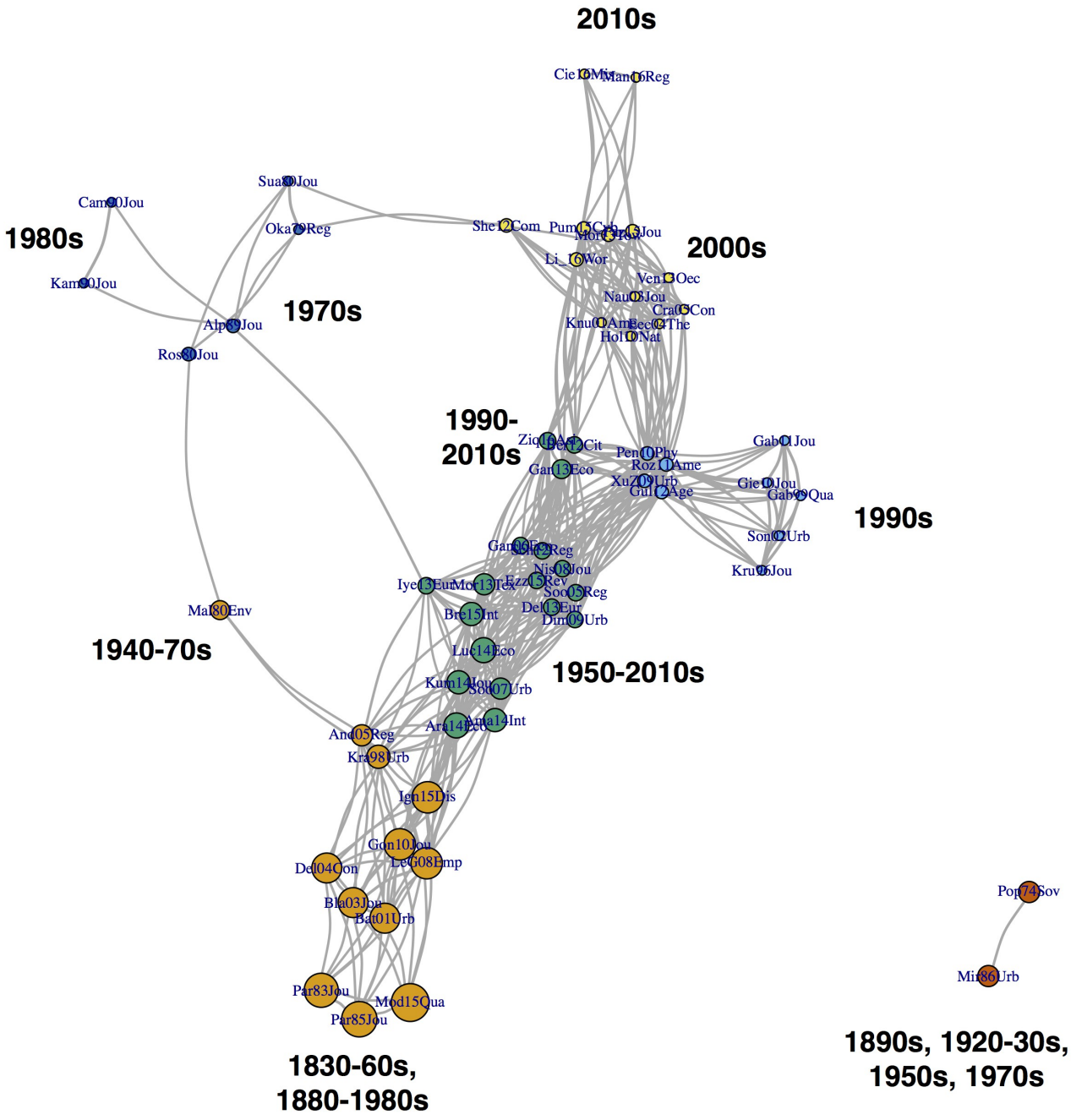


Figure S4. Similarity network of corpus articles by the common city definitions they reported alpha on (cut-off at 0.1).

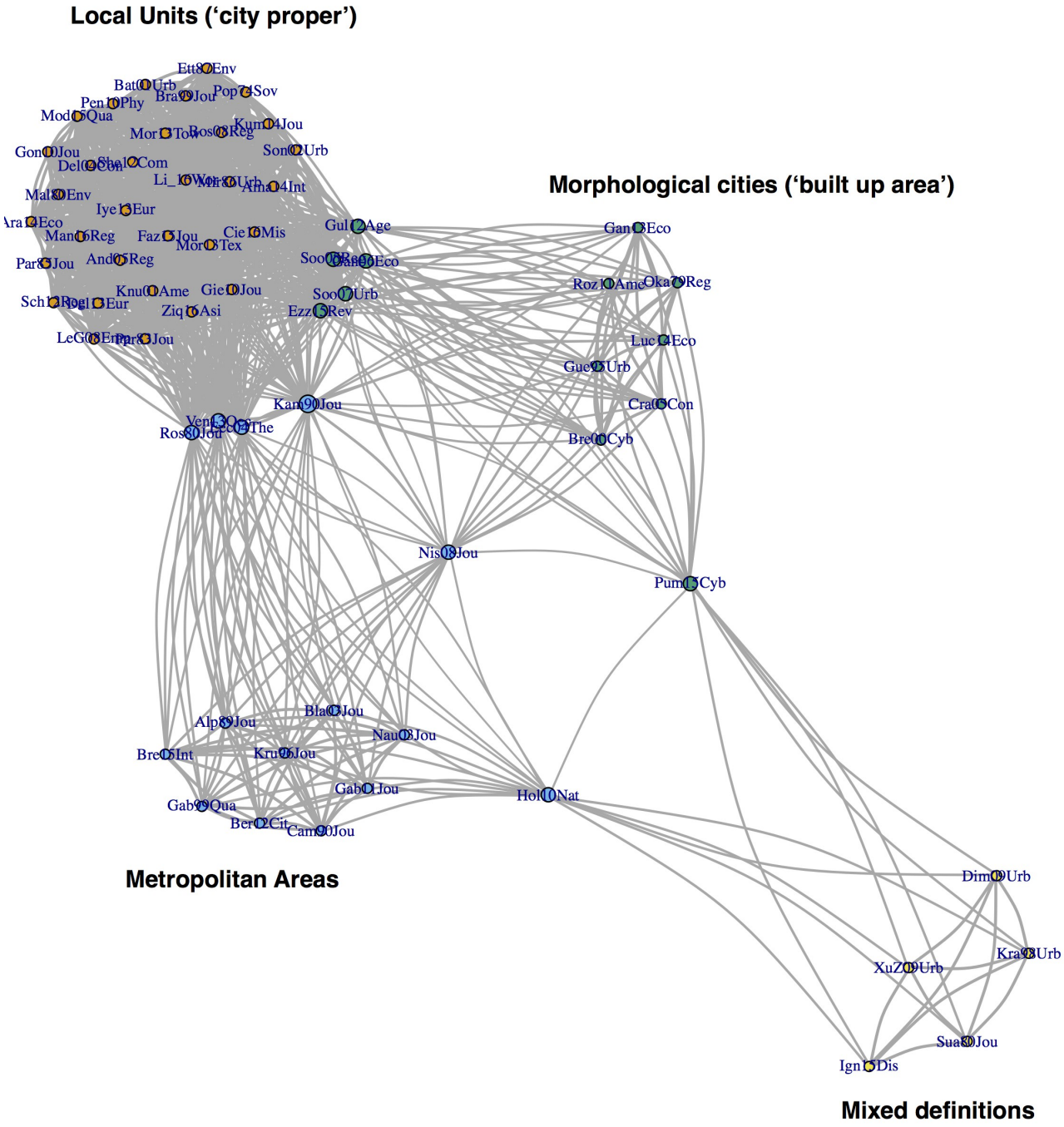
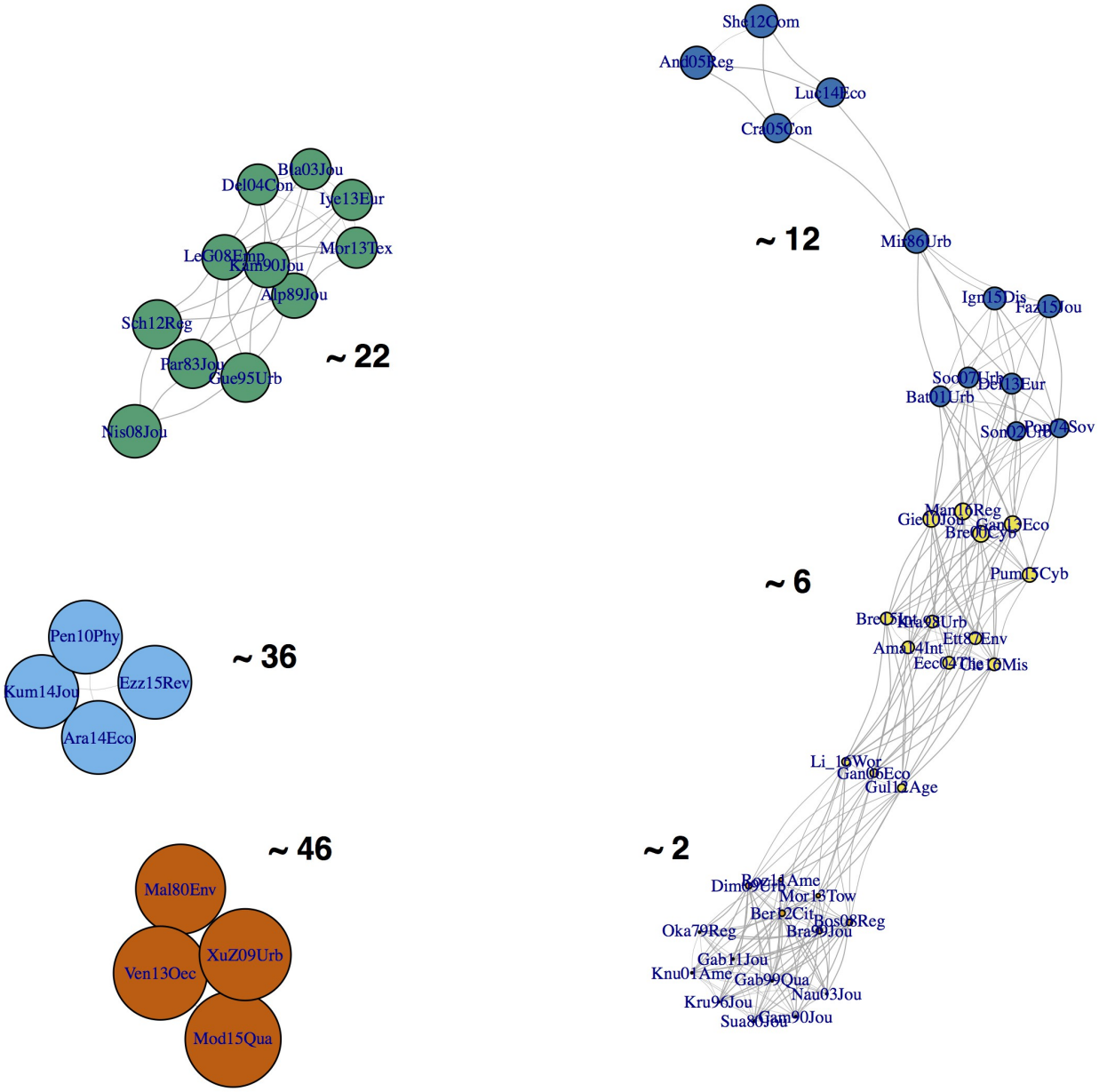


Figure S3. Similarity network of corpus articles by number of estimates reported (cut-off at -0.1).



* the size of nodes reflects the number of estimates reported in the article and the numbers in black correspond to the average number of estimate per articles reported for the community.

Figure S4. Residuals of the model of similarity in mean alpha. Left: Most positive residuals (over 1.1). Right: Most negative residuals (under -2).

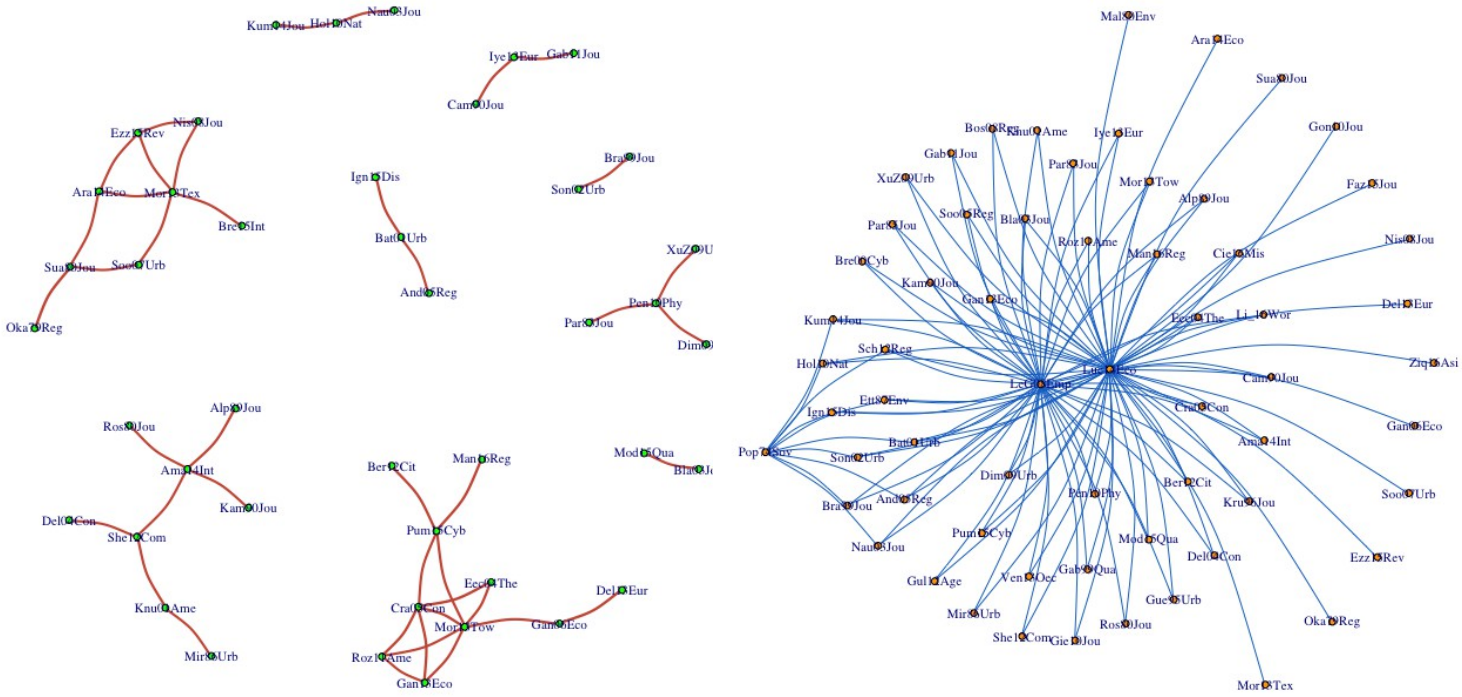


Figure S5. Residuals of the model of similarity in alpha dispersion. Left: Most positive residuals (over 1). Right: Most negative residuals (under -1.5).