Bibliographic Search with Mark-and-Recapture

Chuan Wen, Loe, Henrik Jeldtoft Jensen

Department of Mathematics and Complexity & Networks Group
Imperial College London, London, SW7 2AZ, UK

Abstract

Mark-and-Recapture is a methodology from Population Biology to estimate the population of a species without counting every individual. This is done by multiple samplings of the species using traps and discounting the instances that were caught repeated. In this paper we show that this methodology is applicable for bibliographic analysis as it is also not feasible to count all the relevant publications of a research topic. In addition this estimation also allows us to propose a stopping rule for researchers to decide how far one should extend their search for relevant literature.

Keywords: Bibliographic Analysis, Mark-and-Recapture

1. Introduction

There are many situations where one cannot explicitly count all the instances to determine the size of a population, e.g. the number of polar bears in Western Canadian Arctic [1]. Hence to estimate the population size, a statistical sampling method known as *Mark-and-Recapture* is used in Population Biology [2].

This statistical approximation is not limited to ecology and can be applied to epidemiology [3], linguistics [4] and software engineering [5]. In essence Mark-and-Recapture measures the completeness of a sampling over a set. Hence we applied this methodology to assess the completeness of the bibliography of literature reviews.

A literature review is a summary of a research topic where its source of information is curated by domain experts. The authors often have to rely on

Email addresses: `c.loe11@imperial.ac.uk` (Chuan Wen, Loe), `h.jensen@imperial.ac.uk` (Henrik Jeldtoft Jensen)

specialized search engines like Google Scholar, Microsoft Academic Search, or Web of Science to find all the relevant publications. However the number of results from these search engines can easily be in the order of hundreds of thousands, and most researchers have to rely on their gut feelings to stop their search.

This is a similar problem faced by clinical researchers as the results of medical trials are disparate in different databases (Medline, EMBASE, CINAHL, and EBM reviews). Thus clinical researchers used Mark-and-Recapture as a stopping rule to estimate the completeness of their research [6, 7, 8]. In this paper we extend the idea to different disciplines and assess the quality of academic search engines. Finally we also show that the same mathematics can be used to measure the similarity of truncated-ranking where only the ordering of the top few elements are known.

2. Population Estimation

It is highly probable that the bibliography of the literature reviews are incomplete. Just like population biology, it is not possible to capture all the animals to determine the population of an animal species. Hence *Mark-and-Recapture* can be used to approximate the population by sampling the species repeatedly and discounts for the number of instances that were caught previously.

2.1. Mark-And-Recapture

Animals are captured and marked before releasing them back in the wild. After enough time has passed to allow a thorough mixing, the population is sampled for the second time. In the second sample, the ratio of marked animals (from the first capture) to the number of captured animals is approximately the ratio of captured animals in the first sample to the total population, hence by the Peterson method [2]:

$$\text{Total population} \approx \frac{N_1 N_2}{R}, \tag{1}$$

with standard deviation

$$\sigma = \sqrt{\frac{(N_1+1)(N_2+1)(N_1-R)(N_2-R)}{(R+1)^2(R+2)}}, \tag{2}$$

where N_1 and N_2 are the number of captures in the 1^{st} and 2^{nd} sample respectively, and R is the number of marked animals (individuals that were captured in both samplings). For multiple captures, the weighted variant of Eq. 1 is known as Schnabel Index [9]:

$$\text{Total population} \approx \frac{\sum_{i=1}^{m} N_i M_i}{\sum_{i=1}^{m} R_i}, \tag{3}$$

with standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{m} R_i}{(\sum_{i=1}^{m} N_i M_i)^2}}, \tag{4}$$

where N_i is the number of captures in the i^{th} sample, M_i is the total number of marked animals in the population before the i^{th} sample, and R_i is the number of marked captures in the i^{th} sample.

2.2. Assumptions in the Estimation

To apply the same methods to citation analysis, the assumptions have to be parallel to Population Biology. The mixing period for population biology has to be long enough such that the second sampling is independent from the first, yet short enough to minimize the effects of population changes or the death of the tagged animals, i.e. the system is a closed population. Hence the literature reviews have to be independent efforts and published around the same time.

However the probability that a paper is found and referenced is not equal [10]. There are many factors that affects the visibility of a publication in a search engine (respectively literature review), e.g. quality of research, discipline, keywords, date of publication, authors, etc. This is a common violation of assumption in wildlife as some animals have a higher tendency to be captured again, i.e. "trap-happy" animals. Therefore we can assume the result is the lower bound to the true population size.

The above assumptions are similar for the comparisons of the different independent search engines. The i^{th} (top) article of a search result is analogous to spotting the i^{th} whale in the wild and it is "marked" by cataloging the whale's unique features on its hump. When a whale's unique features are already in the catalog is it then "recapture". This is known as "Sight-and-Resight". In addition the sequential occurrence of the articles/whales allows us to do some time series analysis on the data.

3. Comparing the Bibliography of Literature Reviews

3.1. Experiment Methodology

There are several reviews on the community detection algorithms of graphs over the past decade — Newman 2004 [11], Fortunato and Castellano 2007 [12], Schaeffer 2007 [13], Porter *et al.* 2009 [14], and Fortunato 2010 [15]. Although it is tempting to apply Schnabel Index to sample the body of literature repeatedly, it violates many assumptions of the estimator which will make the results questionable.

The first violation is that these surveys are not independent sampling of the literature as most of them cited the earlier reviews. Secondly the population in question is not closed as there are many publication on communities detection since year 2004. There is only 44 references in Newman 2004 review versus the 457 references in the review by Fortunato in 2010. Thus the results will be meaningless even if the numbers appears to support the methodology.

Therefore to minimize the violation of the assumptions, the reviews must be published approximately the same year and the latter should not cite the earlier review. Hence in this case Schaeffer 2007 will be the first sample and the review by Fortunato and Castellano 2007 will be the second. Finally the result will be compared against the bibliography of the review by Fortunato 2010 to gauge the accuracy of this methodology.

3.2. Results

Out of the 249 references in Schaeffer 2007, only 43 articles are directly relevant to communities detection. Most of the excluded references are on graph cutting from graph theory or clustering algorithms from machine learning as they do not connote the idea of modularity of communities in the articles. Similarly only 55 articles are chosen from the 97 references in the review by Fortunato and Castellano 2007.

Finally since there are only 20 relevant citations that were listed in both reviews, Eq. 1 and Eq. 2 suggest that there are $\approx 118 \pm 14$ publications on graph communities by 2007. In comparison, there are 112 articles before 2008 on graph communities in the bibliography of Fortunato 2010. The agreement is surprisingly good and supports the framework to use Mark-And-Recapture to determine the completeness of a literature review.

4. Comparing Search Engines

Since literature reviews are well curated, the estimate from Mark-And-Recapture may suggest the size of the body of literature on a given topic. It gives new researchers a level of confidence in their preliminary investigations.

However the conditions for this methodology are hard to meet (section 2.2) for most research topics. Furthermore, *is the bibliography of the literature reviews even complete?*. Since academic search engines are the basic sources of information for researchers, we applied Mark-And-Recapture to compare the results from the different search engines.

4.1. Related Work

The preliminary process of a research is the task of searching and *re-searching* the relevant publications to provide a comprehensive overview of a topic. There is no optimal stopping rule to determine if one has collected sufficient relevant articles, especially prolonged search will eventually reach a point of diminishing returns. This is a foremost challenge for any researchers and one of the reasons for peer reviewing publications (i.e. to avoid duplicated research).

The right balance for the time needed to find the relevant materials is of particular interests for medical research. Given the growing amount of research versus the urgency to provide the proper medical care, the research time has to be optimized. However the citation network of related clinical trials is disconnected, which reflects the possibility that the "different camps" of clinical researchers use different research tools and hence are unaware of the relevant literature from the other "camps" [16].

Thus Mark-And-Recapture methodology was proposed as a stopping rule for medical medical research [6, 17, 7, 18, 19]. For example the empirical evaluation on osteoporosis disease management publications estimates approximately 592 articles are missing from 4 main bibliographic databases — MEDLINE, EMBASE, CINAHL, and EBM Reviews [18].

4.2. Experiment Methodology

The above framework however cannot be easily adopted for many fields of science. Many keywords have multiple meanings in different contexts, for example the word *graph* can be defined as a plot of a function or an abstract mathematical object. Hence there can be many unrelated results and thus the search engine can easily return hundreds of thousands of articles.

One way to sieve through the articles is to accept the "top" few relevant articles (suggested by the search engine) until no new significant information is gained [20]. However the measure of *information gain* cannot be quantified and is often based on our subjective gut feelings. In this paper we address this issue by using Mark-And-Recapture on the following academic search engines: Google Scholar, Microsoft Academic Search, and Web of Science.

The web-crawler and the database of these search engines are the "traps" for the entire body of literature, and the ordering of the results is a reflection of the (search engine) algorithms' unique perspectives of the keywords. Suppose the top n^{th} results of two search engines, E_1 and E_2, have R number of common articles. Eq. 1 suggests that there are at least a total of $T = n^2/R$ publications on this topic. To avoid the division by zero, we initialized $R = 1$.

If we assumed that one stops at the n^{th} entry of E_1 and E_2, then the coverage of the body of literature is at most $C = (2n - R)/T$. Therefore the rate of change of C with respect to n estimates the information gained during the time spent with the search engines. A low rate of change implies low information gain and quantifies a stop to the search.

For simplicity, this paper only compares two search engines at a time where each of them is independent samplings over the body of literature. The ordering of the results is sorted by "relevance" which is ranked by the different algorithms of the search engines.

Lastly only the top 500 results from each search engine are collected in the experiments since Web of Science limits that number of articles to be exported at each time. Moreover if the sampling is too large it will trigger Google Scholar to temporarily ban users from accessing its database. The software used to extract from Google Scholar and Microsoft Academic Search is *Publish or Perish* [21].

4.3. The Results from the Comparisons of the Search Engines

Some papers are published in multiple sources, e.g. arXiv and peer-review journals and it will cause the search engines to occasionally return the same paper as multiple and distinct publications. Since there is no information gain for repeated articles, we have to adjust our equations.

The coverage C of a literature is a time series where the n^{th} unit of time refers to the n^{th} article of the search engines. Let $N_{i,n}$ be the number of *unique* articles returned by search engine E_i at time n. If T is the estimated number of publications on this topic at time n, then Eq. 1 gives us:

$$T = N_{1,n}N_{2,n}/R, \tag{5}$$

where R is the number of unique articles that are found in both search engines. Similarly, the coverage of the body of literature is adjusted as:

$$C = (N_{1,n} + N_{2,n} - R)/T. \tag{6}$$

In most cases $N_{1,n} = N_{2,n} \approx n$, which is the easiest to analyze. If R converges to a constant, then $\lim_{n\to\infty} C \approx 1/n \to 0$. This implies that the further you continue the search with the same keywords, there is a diminishing returns to the information gain.

From another perspective if R converges to a constant, then $\lim_{n\to\infty} T \approx n^2 \to \infty$. This implies that the given keyword is so imprecise that the results from the different search engines diverge as there is almost no common articles between the search engines.

In contrast if the rate of growth of R is close to n, then there is at most n common articles at time n. Although the coverage $C \approx 1.0$ and the estimated total number of articles is n, the figures are not meaningful. This is because it implies that the results of E_1 and E_2 are so similar that it is analogous to using only one search engine. In such case we are back to the original situation where there is no quantified method to analyze the results.

Fortunately R generally does not grow in such a way for the entire time series and can be analyzed by plotting T as a function of n. In fact R tends to be sublinear and the coverage will approach zero. Hence the optimal stopping rule is to stop at a point when the derivative of C is zero, where it implies that the search has diminishing returns.

At the local maximum of C, further search have negative returns as the search engines' perspectives of the keyword begin to diverge. This is supported by the quadratic growth of T after the stopping point. Hence the reason to stop is that the subsequent articles are less relevant from the perspective of the other search engines.

At the local minimum of C, the stopping rule is slightly counter-intuitive. As the coverage increases, technically it is prudent to continue the search as it implies that the researcher has more complete coverage of the literature. However for the coverage to increase rapidly, R has to increase rapidly too. It usually means that the subsequent articles are already returned in the earlier results, and hence no information gain.

Finally if $N_{1,n} \neq n$, then T is sublinear. This implies that some literature are published in multiple journals/sources. By definition, two articles are the same if they have the same title and authored by the same researchers.

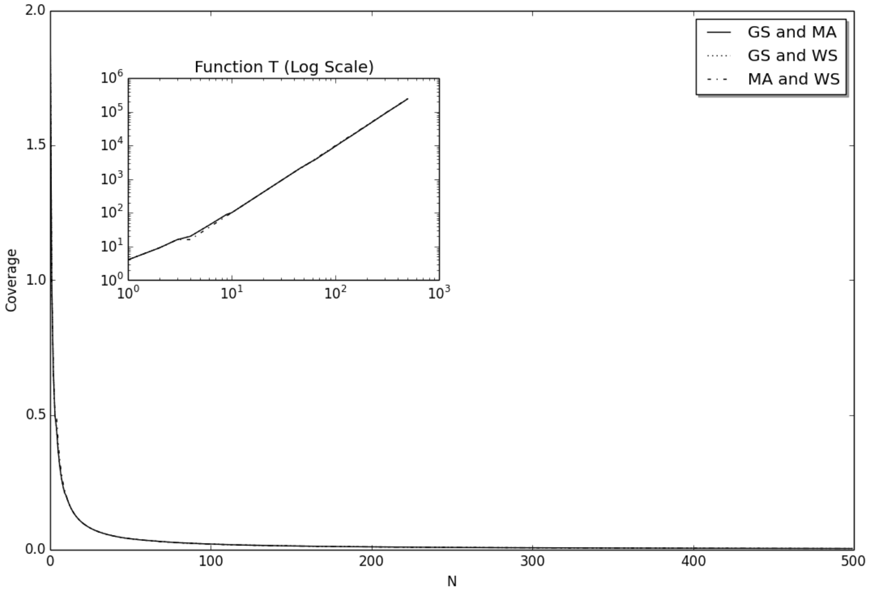

Figure 1: **Keyword: Rechargeable Batteries**. GS, MA and WS are abbreviations for Google Scholar, Microsoft Academic Search and Web of Science respectively. T is quadratic for all pairwise comparisons (R grows so slowly that it is almost constant), hence in the inserted figure it is linear in log scale. As one searches further into the results with such a general keyword, one does not get more focused/specialized in the field and thus the coverage approaches zero for increasing n.

4.4. Empirical Results

The keywords chosen in this paper are primarily based on our familiarity with the topics in Physics and Computer Science. The remaining keywords from the other disciplines are selectively chosen from ScienceWatch.com publication on the top 100 key scientific research front for 2013 [22].

4.4.1. Type I (Convergence to Zero)

The quality of a search depends on how specific the keywords are, for example many disciplines like physics, chemistry and engineering have subfields that research on improving rechargeable batteries. Hence the results from different search engines are drastically different with keywords like "rechargeable batteries" (Fig. 1).

Therefore if a keyword has graphs that is similar to Fig. 1, it suggests

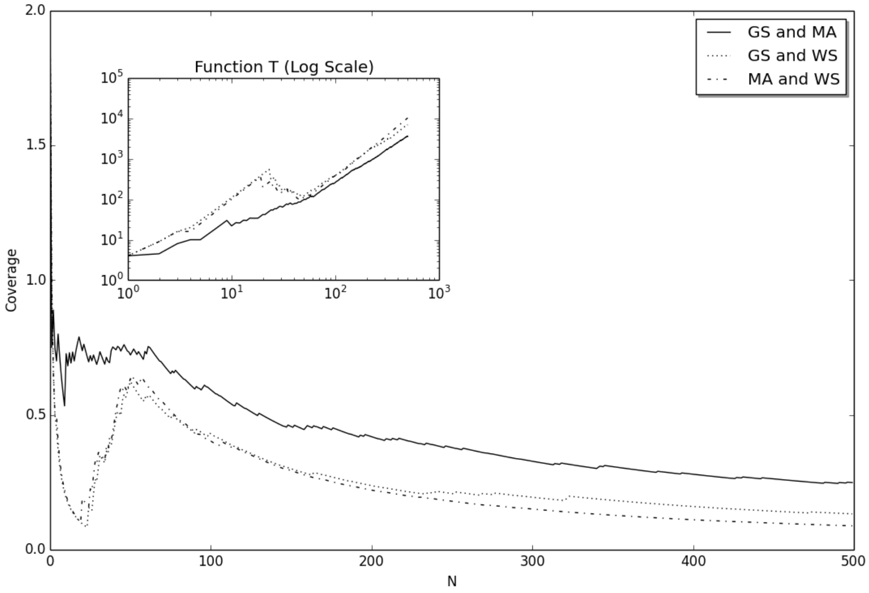

Figure 2: **Keyword: Kauffman Model**. At $n \approx 70$ (local maximum), the rate of change of coverage shifted from zero to negative. This implies one should stop around this point as further search has negative returns. An alternative stopping point is at $n \approx 20$ (local minimum) where it implies that the subsequent articles are already found by the other search engine.

that one should refine the keyword to be more specific. The keyword is either too ambiguous like "Phase Transition" and "Communities Detection", or the topic is studied in many branches of science like "Genetic Algorithm" and "Ising Model". In such cases, there is no good stopping rule.

4.4.2. Type II (1 Local Max and Min)

One way to suggest that the search results are drastically different is when T grows quadratically. This usually implies that the choice of keywords is bad and one should discard the search results. However it is not true in general, for example consider the keyword "Kauffman Model" in Fig. 2.

The local minimum of C (for dotted and dashed line) is approximately at $n = 20$ where T appears to be linear in log-scale (i.e. polynomial growth). The rapid increase of coverage peaked approximately at $n = 50$ is the effect that the subsequent articles after $n = 20$ in one of the search engines were

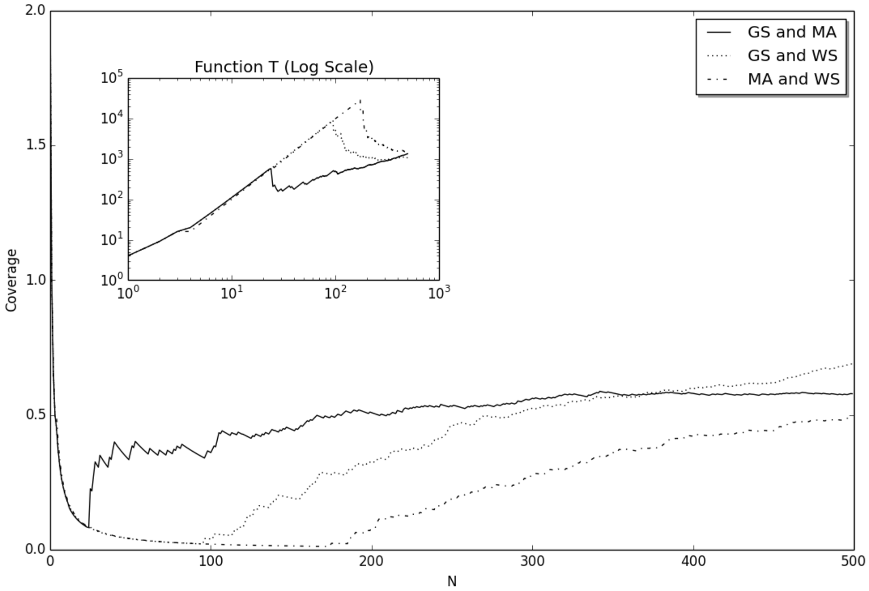

Figure 3: **Keyword: Skyrmion.** In the initial, T grows quadratically, this implies that Web of Science and Google Scholar are significantly different. However at $n \approx 100$, T begins to decrease rapidly and subsequently grows linearly. This implies that the later articles in Web of Science matches the earlier articles in Google Scholar.

already listed in the search result of the other search engine. Thus there is little information gain and it is a reasonable to stop at $n = 20$.

The local maximum of C plateaued until $n \approx 70$, where it is an alternative stopping point for the search. It is an indicator that the search engines' suggestions begin to deviate and hence subsequent articles are less relevant to the keywords. Thus continuing search yields negative returns, which is worse than diminishing returns.

Keywords with graphs that are similar to Fig. 2 are unfortunately not very common. Out of the 50 keywords selected for our experiments, only the graphs of "Kauffman model" and "Tangled Nature Model" have both local minimum and maximum.

4.4.3. Type III (1 Local Min)

There are many examples that fall into this category, especially for keywords that are less ambiguous and found in very specialized topics. For example "Skyrmion" has approximated 9000 articles in Google Scholar and most of the publications are also in the database of the other search engines. However every search engines have their own unique algorithms to rank the most relevant articles.

Fig. 3 shows that the results by the Web of Science initially deviates from Google Scholar and Microsoft Academic Search until $n \approx 100$ and $n \approx 180$ respectively. After which T converges for all pairwise comparisons. This implies that the initial ordering of "relevance" by Web of Science is partially the reverse of the result of Google Scholar.

More precisely after the local minimum, the subsequent articles by Web of Science are found in the earlier results of Google Scholar and Microsoft Academic Search. Therefore the coverage increases and there is little information gained. Thus for example if one uses Google Scholar and Web of Science, one should stop the search at $n \approx 100$ to avoid diminishing returns as the subsequent articles are mostly found much earlier. This is similar to Type II graphs where one stops at the local minimum.

4.4.4. Type IV (No Significant Feature)

There are many instances where the graphs do not fit into any of the above models due to the nature of the search engines. There is no significant minimum or maximum point for one to suggest a meaningful stop to the search. For example the solid line (Google Scholar versus Microsoft Academic Search) in Fig. 4 is the graph for "Causality Measures".

We are not able to deduce a general rule to identify keywords that fall into this category: "Q-Statistics", "Superconductivity", "Ant Colony Optimization", " DNA Methylation", "Renormalization Group" and "Hubbard Model". However it appears that the keywords are very specific and the corresponding publications tend to be published in highly specialized journals/conferences. Thus it is possible that there is insufficient data to support a stop for such keywords.

5. Measure of Truncated-Ranking Similarities

The order of the results from a search engine is often determined by the relevance of the articles. For instance Google's algorithm has roots from Eigenvector Centrality where it ranks the quality of an article via the behavior of "word-of-mouth" recommendations. I.e. high ranking articles are either referred by other high ranking articles or by many independent articles.

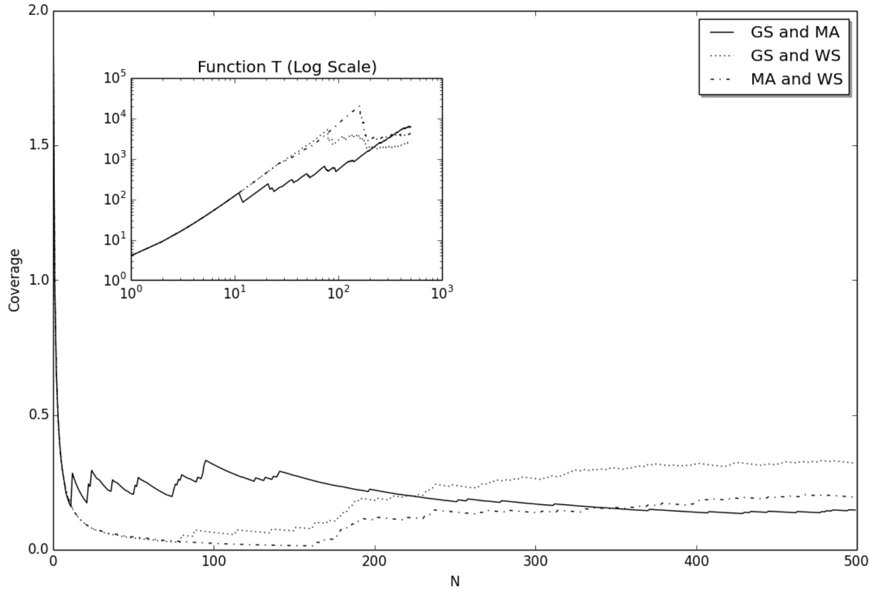

Figure 4: **Keyword: Causality Measures.** There is no significant reference point such that one can suggest a reasonable stop to the search.

Therefore the growth of R in essence is also a measure of similarity for the centrality ranking of vertices (search engines ranking). Specifically a linear R with slope 1 indicates high similarity while slow growing (e.g. sublinear) R indicates a lower degree of similarity. Thus we want to quantify this intuition as a similarity metric between rankings. This is closely related to Spearman's Correlation and Kendall-tau Distance as ways to measure the similarity of ranked variables.

Spearman's Correlation is the variant of Pearson's Correlation for ranked variables where it measures the degree of monotonic relation between two variables. Although it is relevant to our application, the model cannot be used for comparing truncated-rankings, e.g. comparing the top 100 elements of two rankings (on millions of elements). Thus it is also not applicable for dynamical systems where the size of the network fluctuates and only the top centrality vertices are interesting.

Kendall-tau Distance (See Appendix A) measures how likely it is that the order of two rankings agree. It handles truncated-ranking by ignoring element pairs that do not exist in both rankings. It is sensitive to the ordering of the elements and two rankings are independent (dissimilar) if they are random permutation of each other.

It is a good metric until one considers the size of the entire system. It is highly unlikely by random chance in a large system that the top elements of two rankings are in common. Thus even though the orderings of two truncated-rankings might not agree in general, this effect is small relative to the fact that the number of common elements between two truncated-rankings is great.

5.1. Squared Error as a Metric

The intuition of this metric is based on the observation that when two truncated-rankings are identical, R is a straight line with slope 1 intersecting zero (i.e. y=x). However when two truncated-rankings are totally dissimilar, i.e. none of the top vertices in one of the ranking is among the top vertices of the other, R is a straight line with slope 0 (i.e. y=0).

Thus to measure the similarity between two truncated-rankings, we used the Squared Error difference between R and the line $y = x$. The smaller the Squared Error, the more similar two rankings are. If two rankings do not have the same vertices or the ordering of the vertices are different, then the Squared Error will increase and hence indicate lack of similarity. This idea is based on the best fit line algorithm where the Squared Error between the data and line is minimized.

The maximum Squared Error is the difference between the lines $y = x$ and $y = 0$, hence to normalize the measure:

$$S = 1 - \frac{E(I, R)}{E(I, Z)}, \tag{7}$$

where for the top n elements, $I = \{1, 2, \ldots, n\}$ is the ideal case (y=x) and $Z = \{0, \ldots, 0\}$ (n zeros) is the case where there is no similarity. The Squared Error E is defined as:

$$E(X, Y) = \sum_{j=0}^{n} |x_j - y_j|^2. \tag{8}$$

5.2. Experiments Methodology

To simulate a dynamic network that varies in size, we construct a process that adds and removes random vertices from a network in each time step.

Between each iteration, the Eigenvector Centrality of the vertices are computed and only the top 1000 vertices are compared. For example let G_t and G_{t+1} be the networks at time t and $t + 1$ respectively. If Q_t and Q_{t+1} are the ordered lists of the top centrality vertices of G_t and G_{t+1} respectively, then R is derived by comparing Q_t and Q_{t+1} in the same way as we did with search engines in the section 4.

We will begin with a network on 10000 vertices constructed using Barabási-Albert's construction (See Appendix B). In each iteration, x_r random vertices are removed and x_a vertices are added to the network where x_r and x_a (rounded to the nearest integer) drawn from a normal distribution with mean 1000 and standard deviation 100. The new x_a vertices are added into the network using the same mechanism from Barabási-Albert's construction.

To further distinguish the Squared Error metric from the Kendall-tau Distance, we will present some special cases in the experiments to demonstrate their differences. Lastly we will measure the similarity of search engines using the real world data in the previous section.

5.3. Empirical Results

5.3.1. Synthetic Network

Let Q_1 and Q_2 be two truncated-rankings on the index of vertices of a network. From the 1000 iterations in the experiment, the similarity S has a mean of 0.8831 with standard deviation of 0.0697. It is highly correlated to the size of the set $Q_1 \cap Q_2$ with a Pearson's Coefficient of 0.984.

In contrast S is less correlated (Pearson's Coefficient of 0.2443) to the Kendall-tau Distance as there are significant changes to the ordering of the top centrality vertices. More importantly the mean Kendall-tau Distance is 0.0332 with standard deviation of 0.0285. This implies that the Kendall-tau Distance claims that the two truncated-rankings are dissimilar. The main reason for this dissimilarity is that there are many vertex pairs in one ranking that are not in the ranking of the other.

For example let $v_i, v_j \in Q_1$ where v_i is ranked higher than v_j in Q_1. Suppose $v_i \in Q_2$ and $v_j \notin Q_2$, then there is neither agreement nor disagreement between Q_1 and Q_2 on the pair (v_i, v_j). If there are many instances of such pairs, then the Kendall-tau Distance will be close to zero and implies that Q_1 and Q_2 are independent. However considering the size of the system, it would be unlikely to find many common top centrality vertices (e.g. v_i). Thus it is counter-intuitive and peculiar to suggest that the two rankings are not similar.

5.3.2. Special Cases

Since $|Q_1 \cap Q_2|$ is highly correlated to our similarity metric S, it may appear that S is not insightful. Hence this section presents some special cases of Q_1 and Q_2 to further distinguish S from the existing metrics.

Reverse Ranking: When Q_1 is the reverse of Q_2, $|Q_1 \cap Q_2| = 1$ and $S = 0.7492$. It will be particularly strange to state that the two truncated-rankings are identical given that $|Q_1 \cap Q_2| = 1$. Therefore our similarity metric distinguishes itself from the naive approximation of $|Q_1 \cap Q_2|$ by considering the order of the elements in the rankings.

Random Permutation: Suppose Q_1 is a random permutation of Q_2 and as before it will be strange to assume that both truncated-rankings are identical since $|Q_1 \cap Q_2| = 1$. In our simulations on 1000 trials, the mean value of S and Kendall-tau Distance is 0.8993 and -0.0016 respectively. More importantly their Pearson's Correlation Coefficient is 0.9423, thus suggesting that our metric S is similar to Kendall-tau Distance when it comes to measuring the ordering of the elements. Thus it further supports the fact that our metric is more sophisticated than the naive approximation with $|Q_1 \cap Q_2|$.

Asymmetry of Ranking: Unlike the other measures, our metric places more emphasis on the top positions of the truncated-ranking. For example let $Q_1 = \{v_a, v_b, \ldots, v_y, v_z\}$, $Q_2 = \{v_b, v_a, \ldots, v_y, v_z\}$ and $Q_3 = \{v_a, v_b, \ldots, v_z, v_y\}$ where the "..." is identical for all three truncated-rankings. For the other metrics, the similarity between (Q_1, Q_2) is the same as the similarity of (Q_1, Q_3). However our metric shows that (Q_1, Q_2) is less similar than (Q_1, Q_3).

Let $|Q_1 \cap Q_2| = |Q_1|/2 = |Q_2|/2$ where the first halves of Q_1 and Q_2 are random permutations of each other. Thus there is no common element between the second halves of Q_1 and Q_2. From 1000 trials, we computed a mean score of 0.8629 and 0.7523 for S and Kendall-tau Distance respectively. Their Pearson's Correlation Coefficient is 0.9573.

If the situation is reversed, i.e. there is no common element between the first halves of Q_1 and Q_2, and the second halves are random permutation of each other, then the mean score of S and Kendall-tau Distance is 0.3616 and 0.7519 respectively. Since Kendall-tau Distance just counts the number of agreement/disagreement to the element pairs, it does not matter if the missing elements are positioned at the beginning or the end of the ranking. This is different from S as the agreement at the beginning of the rankings has a higher score than the agreement at the end of the rankings.

5.3.3. Real World Data

The observation from our real world data (results from search engine) is similar to the results with the synthetic network in the previous experiments. Specifically our metric is positively correlated to the size of $|Q_1 \cap Q_2|$ with a Pearson's Coefficient of > 0.95 for all pairwise comparisons of the search engines. In addition our metric is almost independent to Kendall-tau Distance with Pearson's Coefficient ≈ -0.1.

However it is the absolute score of the metrics that is particularly interesting for this section. For instance between Google Scholar and Microsoft Academic Search, the mean similarity score (over all the search results in section 4) for $|Q_1 \cap Q_2|$ and Kendall-tau Distance are 0.2799 and 0.0068 respectively. This implies that their results are not similar by those measures. In contrast, our metric has a score of 0.464 with standard deviation of 0.2347.

Since the score is normalized between 0 and 1, suppose we let the arbitrary threshold between similarity and dissimilarity to be 0.5. Thus our metric suggests that there is a huge variance in the similarity of Google Scholar and Microsoft Academic Search. This supports the diverging conclusions from other empirical studies that they are *both* similar and dissimilar in general. Therefore our metric is normalized in a way such that it is good for measuring truncated-rankings like search engines' results.

6. Summary

Mark-and-Recapture is a simple statistical approximation used by Ecologists to estimate the population size of a species. It can also be used in applications where one has partial knowledge of the population. Therefore we proposed using this methodology to assess the completeness of the bibliography of a literature review.

As a proof of concept, we have shown that the approximation is accurate to assess the literature reviews on "Communities Detection of Networks". The estimated number derived using the bibliographies from two literature reviews in 2007 is close to the number of relevant articles (prior to 2008) in the bibliography of a highly cited review paper by Fortunato in 2010.

The concept of measuring the completeness of a bibliography is similar to estimating the proportion of relevant articles found for a given topic. If we assume that the authors of these literature reviews used academic search engines to collect their sources, then it will be useful to assess the completeness of the results returned by the search engines. Thus we reapplied

Mark-and-Recapture to study this problem.

The problem has been formulated as a time series (on variable n) where the first n articles are used to obtain the ratio of the literature found by the search engines to the estimated size of the complete literature. This ratio is known as the coverage of the literature and it is a way to measure the fraction of information known at time n. Thus the change of the coverage at time n measures the information gain (or loss) if one is to include the n^{th} article in the research.

Therefore we are able to develop a quantitative stopping criteria for one to follow to maximize his time and resources with the search engine. Lastly the time series also signal the quality of the choice of keywords used in the search engines. It assumes that the search engines are able to pick the most relevant articles of a given topic and if opinions of these search engines fail to converge, then it indicates that one should refine the choice of keywords.

The stopping rules however does not factor the external costs involved in searches, e.g. the effort to organize and digest a huge collection of materials. Thus future work is to address this issue. This could also potentially allow us to quantitatively measure the efficiency of using multiple search engines versus using a single search engine with different keywords.

Finally we show that the same mathematics and ideas can be used to measure the similarity of data-truncated rankings since the problem is parallel to comparing the top articles of search engines. It addresses the issue of truncated ranking in existing similarity metrics like Spearman's Correlation and Kendall-tau Distance. Specifically our metric considers that in a large system, it is unlikely that there are many common elements found in two different rankings.

In addition in our experiments we showed that the metric is more sophisticated than the cardinality of the intersecting set of two rankings. Not only will the metric penalize the disagreement of the ordering of the rankings, it places more emphasis on the ordering of the top ranks.

A quantitative understanding of the behavior of search and ranking allows us to have a more systematic manner to approach, say, a literature search done for research purposes. Mark-and-Recapture is an approximation to how complete a research search is by consolidating the efforts and insights from different sources like literature reviews. However since search engines are now the main source of information, we believed that it will be extremely useful to introduce stopping rules and similarity metrics to study the results from search engines.

Appendix A. Kendall-tau Distance

Given two rankings of ordered sets $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$, a set of n observation is $(x_1, y_1), \ldots, (x_n, y_n)$. A pair of observations (x_i, y_i) and (x_j, y_j) are in agreement if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. The pair is in disagreement if $x_i > x_j$ and $y_i < y_j$ or if both $x_i < x_j$ and $y_i > y_j$. Hence the Kendall-tau Distance is:

$$\tau = \frac{(\text{no. of agreement pairs}) - (\text{no. of disagreement pairs})}{n(n-1)/2}. \tag{A.1}$$

Appendix B. Barabási-Albert Network

Barabási-Albert network [23] is parameterized by m to refer to the number of new edges at each iteration. The network construction begins with some arbitrary small number of vertices connected randomly.

At each iteration, one new vertex of degree m is added. The edges of the new vertex are connected probabilistically with a probability proportional to the degree of the existing vertices. Define $deg(v_i)$ as the degree of vertex v_i. The probability that the new vertex is connected to vertex v_i is given by:

$$p_i = \frac{deg(v_i)}{\sum_j deg(v_j)}. \tag{B.1}$$

This is referred to as preferential attachment.

[1] D. P. DeMaster, M. C. Kingsley, I. Stirling, A multiple mark and recapture estimate applied to polar bears, Canadian Journal of Zoology 58 (4) (1980) 633–638.

[2] T. Southwood, P. Henderson, Ecological Methods, Wiley, 2009. URL http://books.google.co.uk/books?id=HVFdir3qhxwC

[3] A. Chao, P. Tsay, S.-H. Lin, W.-Y. Shau, D.-Y. Chao, The applications of capture-recapture models to epidemiological data, Statistics in medicine 20 (20) (2001) 3123–3157.

[4] J. C. O. Alcoy, The schnabel method: An ecological approach to productive vocabulary size estimation, International Proceedings of Economics Development & Research 68.

[5] A. Chao, M. C. Yang, Stopping rules and estimation for recapture debugging with unequal failure rates, Biometrika 80 (1) (1993) 193–201.

[6] D. Lane, J. Dykeman, M. Ferri, C. H. Goldsmith, H. T. Stelfox, Capture-mark-recapture as a tool for estimating the number of articles available for systematic reviews in critical care medicine, Journal of critical care 28 (4) (2013) 469–475.

[7] M. Kastner, S. E. Straus, K. McKibbon, C. H. Goldsmith, The capture–mark–recapture technique can be used as a stopping rule when searching in systematic reviews, Journal of clinical epidemiology 62 (2) (2009) 149–157.

[8] H. T. Stelfox, G. Foster, D. Niven, A. W. Kirkpatrick, C. H. Goldsmith, Capture-mark-recapture to estimate the number of missed articles for systematic reviews in surgery, American Journal of Surgery 206 (3) (2013) 439–440. doi:10.1016/j.amjsurg.2012.11.017.
URL http://dx.doi.org/10.1016/j.amjsurg.2012.11.017

[9] Z. E. Schnabel, The estimation of total fish population of a lake, American Mathematical Monthly (1938) 348–352.

[10] D. A. Bennett, N. K. Latham, C. Stretton, C. S. Anderson, Capture-recapture is a potentially useful method for assessing publication bias, Journal of clinical epidemiology 57 (4) (2004) 349–357.

[11] M. E. Newman, Detecting community structure in networks, The European Physical Journal B-Condensed Matter and Complex Systems 38 (2) (2004) 321–330.

[12] S. Fortunato, C. Castellano, Community Structure in Graphs, eprint arXiv: 0712.2716.

[13] S. E. Schaeffer, Graph clustering, Computer Science Review 1 (1) (2007) 27–64.

[14] M. A. Porter, J.-P. Onnela, P. J. Mucha, Communities in networks, Notices of the AMS 56 (9) (2009) 1082–1097.

[15] S. Fortunato, Community detection in graphs, Physics Reports 486 (3) (2010) 75–174.

[16] K. A. Robinson, A. G. Dunn, G. Tsafnat, P. Glasziou, Citation networks of related trials are often disconnected: implications for bidirectional citation searches, Journal of clinical epidemiology.

[17] H. T. Stelfox, G. Foster, D. Niven, A. W. Kirkpatrick, C. H. Goldsmith, Capture-mark-recapture to estimate the number of missed articles for systematic reviews in surgery, The American Journal of Surgery 206 (3) (2013) 439–440.

[18] M. Kastner, S. Straus, C. H. Goldsmith, Estimating the horizon of articles to decide when to stop searching in systematic reviews: an example using a systematic review of rcts evaluating osteoporosis clinical decision support tools, in: AMIA Annual Symposium Proceedings, Vol. 2007, American Medical Informatics Association, 2007, p. 389.

[19] A. Booth, How much searching is enough? comprehensive versus optimal retrieval for technology assessments, International journal of technology assessment in health care 26 (04) (2010) 431–435.

[20] G. J. Browne, M. G. Pitts, J. C. Wetherbe, Cognitive stopping rules for terminating information search in online tasks, MIS quarterly (2007) 89–104.

[21] A. Harzing, Publish or perish.
URL `http://www.harzing.com/pop.htm`

[22] C. King, D. A. Pendlebury, research fronts 2013 (2013).

[23] A. L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.