Computer Science on the Move: Inferring Migration Regularities from the Web via Compressed Label Propagation

Fabian Hadiji, Martin Mladenov
TU Dortmund University
Dortmund, Germany
first.last@cs.tu-dortmund.de

Christian Bauckhage
Fraunhofer IAIS
Sankt Augustin, Germany
first.last@iais.fraunhofer.de

Kristian Kersting
TU Dortmund University
Dortmund, Germany
first.last@cs.tu-dortmund.de

Abstract

Many collective human activities have been shown to exhibit universal patterns. However, the possibility of regularities underlying researcher migration in computer science (CS) has barely been explored at global scale. To a large extend, this is due to official and commercial records being restricted, incompatible between countries, and especially not registered across researchers. We overcome these limitations by building our own, transnational, large-scale dataset inferred from publicly available information on the Web. Essentially, we use Label Propagation (LP) to infer missing geo-tags of author-paper-pairs retrieved from online bibliographies. On this dataset, we then find statistical regularities that explain how researchers in CS move from one place to another. However, although vanilla LP is simple and has been remarkably successful, its run time can suffer from unexploited symmetries of the underlying graph. Consequently, we introduce compressed LP (CLP) that exploits these symmetries to reduce the dimensions of the matrix inverted by LP to obtain optimal labeling scores. We prove that CLP reaches identical labeling scores as LP, while often being significantly faster with lower memory usage.

1 Introduction

Many collective human activities have been shown to exhibit universal patterns. However, the possibility of strong regularities underlying computer science (CS) researcher migration has barely been explored at global scale. Fortunately, in the post-Internet era the WWW stores tons of data on researchers which is frequently updated, and we here demonstrate that this information can be utilized to extract migration behavior of researchers and to learn models for the underlying process. However, there are no datasets available on the Web that immediately allow such an analysis. We first need to build a migration dataset to conduct a large-scale investigation of migration. To this aim, we harvested data from different information sources freely accessible on the Web and merged these into bibliographic databases augmented with geo-tags. However, not all information is available — it might actually be impractical to gather — and it is uncertain. Therefore, we have to rely on an AI algorithm to fill in the blank spots. More precisely, we provide a relational view on *Label Propagation (LP)* [Zhu *et al.*, 2003; Bengio *et al.*, 2006] and introduce a novel way to significantly speed it up based on equitable partitions. We call the resulting algorithm *Compressed Label Propagation (CLP)* because the original LP-graph is "lifted" or rather "compressed" before running vanilla LP on the smaller graph. Running CLP results in the first translational dataset for more than a million computer scientists on which we then learn statistical migration models explaining the results in sociologically plausible ways. To verify the quality of our inferred geo-tags and statistical models, we additionally run CLP on an orders-of-magnitude smaller but manually curated dataset. This demonstrates that missing geo-information can be inferred automatically and in turn, statistical patterns of CS researchers migration can be harvested from the Web.

Indeed, people may have had the suspicion that migration follows certain patterns but our results show that it goes beyond folklore. This objectification is important, given that migration and the demographic change is attracting much attention in the media worldwide nowadays. Unveiling, explaining and ultimately predicting these processes remain key challenges in understanding the behavior of science and scientist all around the globe. The statistics belong to the key inputs to policy formulation and funding in research. So far however, data is mostly collected on a national scale and/or access restricted and especially is not registered across researchers. This is surprising, since it is a truly international phenomenon that should be analyzed on a transnational scale: (computer) science thrives on the free exchange of findings and methods, and ultimately of the researchers themselves.

Specifically, we found the following patterns: **(R1)** A specific researcher's propensity to migrate, that means to make the next move, follows a log-normal distribution. That is, researchers are generally not "memoryless" but have to care greatly about their next move. This is plausible due to the dominating early career researchers with non-permanent positions. This regularity of timing events is remarkably stable and similar within different continents across the globe. **(R2)** The propensity to make $k > 1$ migrations follows a gamma distribution, suggesting that migration at later career stages is "memoryless". That is, researchers have to care less about

their next move, since the majority of positions are permanent in later career stages. **(R3)** Consequently, brain circulation, i.e., the time until a researcher returns to the country of their first publication, also follows a gamma distribution. That is, returning is also memoryless. Researchers cannot plan to return but rather have to pick up opportunities as they arrive.

These remarkably simple but strong patterns are somewhat surprising. Reasons to migrate are manifold and complex: political stability and freedom of science, family influences such as long distance relationships and oversea relatives, and personal preferences such as exploration, climate, improved career, better working conditions, among others. However, recall that we estimate distributions from massive data inferred from the Web; this allows one to distill the patterns.

All our contributions, the data, using AI to deal with uncertainty and missing information, compressed LP, as well as the discovered regularities are novel and go far beyond studies typically carried out for migration. So far, studies within sociology were small scale with a few thousand researchers at a national scale or publicly undisclosed. Indeed, extracting large-scale sociological information from the Web has attracted a lot interest. However, most of the work has focused on services like Twitter or Facebook, see e.g. [Burke *et al.*, 2013; Park *et al.*, 2013], and most importantly work directly on the raw data; no AI technique has been used to fill in missing data. Moreover, these information sources are not open access and, hence, results are not reproducible. Closer to our work are studies focusing on migration and mobility data in general but not on computer scientists. For example, [Zagheni and Weber, 2012] have recently analyzed a large-scale e-mail dataset to estimate international migration rates, but not specific to computer scientists, since the occupation was unobserved. Moreover, Zagheni and Weber have not presented any statistical regularities nor dealt with missing information. Using a large-scale, IP-address-based dataset, [State *et al.*, 2013] also investigated mobility data and migration flows. While State *et al*. present a model for migration probabilities between countries, they also used access restricted data and do not aim at unveiling the sociology of CS. Several other people have also looked into migration, e.g. [Stillwell, 2005; Rees *et al.*, 2009], but have considered small scale data only and have also not investigated computer scientists. In contrast, we present the first large-scale migration study for CS inferred from publicly available data using AI, describe the data harvesting process in detail, and report statistical regularities in this dataset. The results actually suggest that Steward's Poisson-log-normal model [Stewart, 1994] for bibliometric/scientometric distributions of productivity can be also used for migration and scales to a transnational, massive level[1]. This complements many other human activities that have been shown to exhibit patterns, see e.g. [Zipf, 1946; Mantegna and Stanley, 1995; Cohen *et al.*, 2008; Gonzales *et al.*, 2008] among others.

Finally, the algorithmic challenge of creating such a large-scale dataset also called for an efficient inference approach. Compressed LP as introduced here is akin to what is known as lifted probabilistic inference [Kersting, 2012]. While CLP is based on the power method, one can also implement LP via *Gaussian BP (GaBP)* and in turn use lifted GaBP to exploit symmetries in LP [Neumann *et al.*, 2011]. However, this does not allow one to use out-of-the-box GaBP implementations since changes to the GaBP algorithm are required to account for the lifted model [Ahmadi *et al.*, 2011]. Moreover, this "lifted" LP approach is based on matrix inversion and requires several re-liftings which is impractical for graphs at massive scale. Instead, we here extend the recent lifting by reparameterization paradigm [Mladenov *et al.*, 2014] to the LP problem, which allows one to lift/compress the LP weight matrix only once. Another recent approach to speed up LP was presented by Fujiwara and Irie [2014], who reduce the run time by updating only the scores of a subset of labels in each iteration. Also similar in spirit to CLP are the ideas of Alexandrescu and Kirchhoff [2007]. They proposed to merge identically labeled nodes to speed up LP, whereas CLP intuitively clusters the entire graph. Actually, there are many more efficient LP approaches, see e.g. the references in [Fujiwara and Irie, 2014] for an overview, and any of them can be used on top of our compressed graph.

We proceed as follows. First, we describe how we harvested the data and discuss how we inferred the missing data from the Web with our logic-rule based LP. We then explain our compressed LP and prove its correctness. We support our theoretical analysis by experiments on two different datasets that show how CLP reduces run times and enables the labeling of larger datasets then before without sacrificing quality. Before concluding, we present the patterns found in the data.

2 Labeling Bibliographies with Geo Tags

The WWW provides several freely accessible bibliographies with millions of papers and authors. However, most of them do not contain affiliations or geo-information. For an extensive study of researcher's migration behavior this information is crucial though. Our goal is to label every author-paper-pair in a bibliography with the affiliation of that author and its geographic location. Although it is possible to manually, or semi-automatically, retrieve such labels, a full labeling of large databases, such as DBLP[2], is not practical by such methods. In addition, if we can build an effective automated machinery that helps us with this task, it is also much easier to update the database continuously with new papers arriving.

To this end, we assume an initial bibliography consisting of papers and theirs authors. We start by adding affiliation information to authors in our bibliographic database. One of the resources that contain affiliation information is the ACM Digital Library[3] (ACM DL). Unfortunately, ACM DL does not allow a full download of the data. Consequently, we retrieved the affiliation information of only a few author-paper-pairs randomly selected from ACM DL which we then matched with our bibliography. This gave us initial seed affiliations per author for different papers. In order to fill in the missing information, we resorted to AI. To do so, however, we have to

[1] Due to space restrictions, we do not present the full migration model but rather focus on data compilation and the patterns found.

[2] dblp.uni-trier.de

[3] dl.acm.org

be a little bit more careful. First, the names of the affiliations in ACM DL are not in canonical form which results in a very large set of affiliation candidates. Secondly, although we have now partial affiliation information, we still lack exact geo-information of the organizations to identify cities, countries, and continents. Many of the affiliation names may contain a reference to the city or country but these pieces of information are not trivial to extract from the raw strings. Additionally, we are interested in latitude and longitude values to enable further analysis and visualization. Hence, we used Google's Geocoding API[4] to resolve the locations. This resulted in geo-tags for most of the affiliation strings. A remaining gap rises from the fact that the API does not find geo-locations for all the strings in our database. Essentially, this is because the strings contain information not related to the geo-location such as departments, e-mail addresses, among others.

3 Relational Label Propagation

Before we infer the missing author-paper-pairs, we revise our obtained affiliations. To further increase the quality of our harvested data, we hypothesized that there are actually not that many relevant organizations as obtained from ACM and these names need to get de-duplicated. Since we have geo-locations for most of the affiliations, we can use this information for a simple entity resolution and cluster affiliations together for which the retrieved cities coincide[5].

Based on these seed geo-locations, we filled in the missing ones using LP. LP runs on an undirected graph $G = (V, E)$ where V is a set of nodes and E is set of weighted edges. Each node corresponds to an author-paper-pair in our bibliography. The edges represent the similarity between nodes. In the following, we use logic rules to formulate this similarity. These rules are based on relations such as co-authorship between the authors associated with the nodes. Specifically, in order to define the edges, we considered the following functions over nodes that return facts about the nodes corresponding to the function name: $\texttt{author}(i)$, $\texttt{paper}(i)$, and $\texttt{year}(i)$. For shorthand, we write $\texttt{a}(i)$, $\texttt{p}(i)$, and $\texttt{y}(i)$. Based on these functions, we defined rules that add a weight λ_k to the each edge weight w_{ij} whenever the rule holds. Initially, we set all weights w_{ij} to zero. The first rule, $w_{ij} = w_{ij} + \lambda_1$ if $\texttt{p}(i) = \texttt{p}(j)$, adds a weight between two nodes if the nodes belong to two authors that co-author the paper associated with nodes i and j. The second rule,

$$w_{ij} = w_{ij} + \lambda_2 \text{ if } \texttt{a}(i) = \texttt{a}(j) \wedge \texttt{y}(i) = \texttt{y}(j)$$

adds a weight whenever two nodes corresponds to different publications by the same author in the same year. Finally,

$$w_{ij} = w_{ij} + \lambda_3 \text{ if } \texttt{a}(i) = \texttt{a}(j) \wedge \texttt{y}(i) = \texttt{y}(j) + 1$$

fires when the nodes belong to two publications of the same author but written in subsequent years. Using these edge weights, we construct an *affinity matrix* $W \in \mathbb{R}^{n \times n}$. If

[4]developers.google.com/maps/documentation/geocoding

[5]Indeed, this approach does not distinguish multiple affiliations per city such as *MIT* and *Harvard*. However, it is simple and effective, and — as our empirical results show — the resolution is sufficient to establish strong regularities in the timing events.

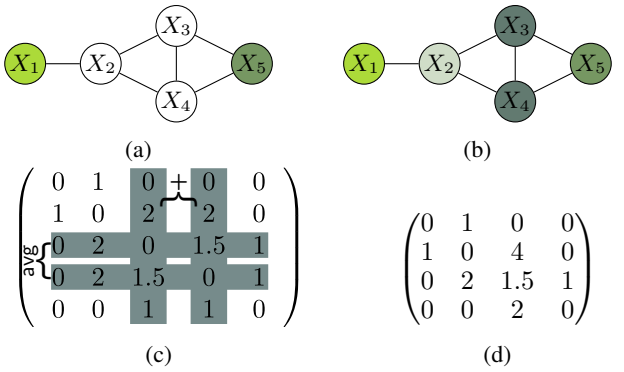


Figure 1: Toy example for CLP. The LP graph in Fig. 1a contains two labeled nodes (green shades) and three unlabeled nodes (white). The corresponding similarity matrix is given in Fig. 1c. Based on the CEP, the graph can be colored as in Fig. 1b. The partition clusters X_3 and X_4 together. The corresponding compressed matrix Q is depicted in Fig. 1d. (Best viewed in color)

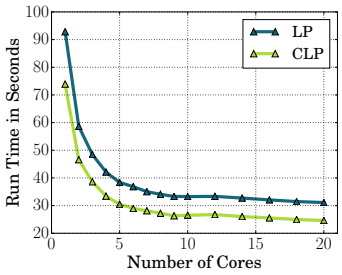
$T = D^{-1}W$ with $D_{ii} = \sum_j w_{ij}$, we can implement LP using a simple power method: $Y^{t+1} = TY^t$, where Y^t is the *labels matrix*. At convergence, row i in Y^* corresponds to a distribution over the possible labels for node i. In Y^0, we set a cell y_{ij} to 1 if we know that node i has label j. All other cells are set to 0. The original implementation suggests a push-back phase in every iteration, clamping the rows of the known nodes in Y^{t+1} to their original distribution as in Y^0. Instead, we adapt the affinity matrix in such a way that we do not need the explicit push-back anymore. More precisely, for a labeled node i, we set $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ otherwise. This iterative procedure is performed until convergence or a maximum number of iterations has been reached. At convergence, the labels of the unknown nodes are read off the labels matrix, i.e., the label of node i is given by $y_i^* = \arg\max_j y_{ij}$.

4 Compressed Label Propagation

While W, and respectively T, is very sparse, Y^t becomes denser with every iteration. Eventually, this presents an obstacle in terms of both computation time and memory requirements. To alleviate some of this burden, we can exploit the latent symmetries in the structure of T. In our proposed approach, CLP, we do so by means of equitable partitions.

The algorithm proceeds as follows (illustrated in Fig. 1): we first partition the nodes according to their initial labels (Fig. 1a). We then compute the Coarsest Equitable Partition (CEP) of T which preserves the initial label partition. From the partition, we obtain a (hopefully) smaller quotient matrix Q by: a) replacing the set of all columns corresponding to nodes in the same class by their sum; b) replacing the set of all rows of nodes in the same class by their average (Fig. 1c). We carry out step b) on Y^0 as well to obtain the compressed label matrix J^0. Finally, we run LP with Q and J^0 (Fig 1d) in place of T and Y^0. As we will show now, we can perfectly recover Y^k from J^k and, thus, the result of LP can be recovered from

Labeled Nodes	Accuracy
6%	0.58
12%	0.67
18%	0.72
24%	0.75
30%	0.78
36%	0.80
42%	0.81
48%	0.82
55%	0.83



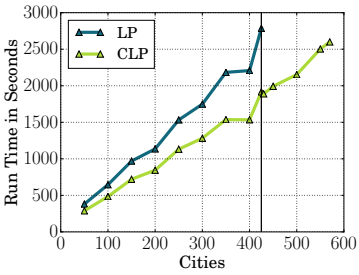


Figure 2: (left) Label accuracy for the AAN dataset with a varying number of initially labeled nodes. One should note that accuracy is a very challenging performance measure for a multi-label problem with around 800 classes. (middle) Runtime of LP on the AAN dataset with an increasing number of available cores. (right) Due to the large size of DBLP, it is not possible to run LP with all cities in a single run. Instead, we have to split the cities into batches and run (C)LP on each batch separately. As one can see, CLP can use up to 570 cities in a single run while LP can only handle 425 cities at once. (Best viewed in color)

the result of CLP.

Theorem 1 (Compressed LP is sound and complete). *Running the power method on the compressed matrix Q returns identical label scores as running LP on T.*

Proof (sketch). First observe that algebraically, Q and J^0 can be written as $Q = \widehat{B}TB$ and $J^0 = \widehat{B}Y^0$, where B is an $n \times p$ matrix having $B_{ik} = 1$ if node i is in class k and 0 otherwise (representing the summing of columns). $\widehat{B}$ is defined as $\widehat{B}_{ki} = 1/|\text{class } k|$ if i is in class k, otherwise 0 and represents the averaging of rows. We first reference the following facts [Grohe *et al.*, 2014]: (♣) $\widehat{B}B = I_p$ ($\widehat{B}$ acts as left inverse of B); (♡) $B\widehat{B}T = TB\widehat{B}$ (the matrix $B\widehat{B}$ commutes with T since B comes from the CEP of T).

As a first step, we need to show that $Y^{k+1} = B\widehat{B}Y^{k+1}$. We proceed by induction. Due to space constraints we omit the discussion of $k = 0$, which follows from the construction of the CEP. For the induction step, we have

$$Y^{k+1} = TY^k \overset{\text{ind.}}{=} TB\widehat{B}Y^k \overset{\heartsuit}{=} B\widehat{B}TY^k = B\widehat{B}Y^{k+1} \ , \quad (\spadesuit)$$

where the second equality follows from our induction hypothesis. Note that we also omitted the discussion of the push-back operation, however, it can be shown that the above holds after push-back as well. Finally, induction shows that $Y^{k+1} = BJ^{k+1}$:

$$\begin{aligned} BJ^{k+1} &= B(\widehat{B}TB)J^k \overset{\text{ind.}}{=} B\widehat{B}T(B\widehat{B})Y^k \overset{\heartsuit,\clubsuit}{=} B\widehat{B}TY^k \\ &= B\widehat{B}Y^{k+1} \overset{\spadesuit}{=} Y^{k+1} \ . \end{aligned}$$ □

Observe now that $Q \in \mathbb{R}^{p \times p}$, where p is the number of classes of the CEP of T. That is, we have one row and column per cluster instead of per node. Thus, if $p \ll n$, we need to solve a much smaller system. Moreover, using the highly efficient implementation of SAUCY [Katebi *et al.*, 2012], CEP computation is done in $\mathcal{O}[(m+n)\log n]$ time; even in case of little to no symmetry, there is only little computational overhead due to symmetry detection.

5 Inferred Regularities from Bibliographies

With CLP at hand, let us now turn towards inferring regularities. There are different choices as a starting point for the data harvesting process. Ultimately, we are interested in a bibliography covering all different scientific disciplines. However, to begin with, we focus on CS. For an qualitative evaluation, we are interested in a dataset with as much ground truth as possible. On the other hand, for an in-depth analysis of researcher's migration behavior, we would like to construct a database as large as possible. Resulting from these different requirements, we will evaluate our CLP on two different datasets to show its efficiency and effectiveness. In particular, we demonstrate (1) that relational LP produces meaningful label distributions with high accuracy on a manually curated dataset and that (2) CLP significantly speeds up standard LP and requires less memory at the same time, which is especially important for large-scale datasets. For all our experiments we used $\lambda_1 = 1$, $\lambda_2 = 3$, and $\lambda_3 = 2$ (found by a grid search on a small subset of the data) as weights for the logical rules described above. LP heavily relies on an efficient implementation for multiplying a sparse matrix with a dense matrix. To this end, we used LAMA[6], a very efficient parallelized C++ linear algebra library. All experiments were run on a Linux machine with 64GB RAM and 20 cores.

5.1 Empirical Investigation of Compressed LP

To verify the quality of relational LP, we start our analysis with a dataset for which we have a relatively large amount of affiliations in advance. The **AAN** dataset [Radev *et al.*, 2009] contains 19,410 publications in total written by 15,397 authors. After reducing the available affiliations to the city level, the resulting number of author-paper-pairs is 49,530 while 33,061 of these nodes are labeled with one of 802 cities. The graph G has a total of 145,594 edges, resulting in a very sparse matrix T. By removing an increasingly number of labels from the graph, we construct test sets of different sizes which we use for the evaluation. We start by removing 10% of the labels, obtaining a graph with 55% of the nodes labeled. We then gradually add 10% of the nodes to the test set until only 6% of the nodes are labeled. We apply this dataset construction ten times, to allow for multiple re-runs of the experiment. The table in Fig. 2(left) shows the average accuracy of the predicted labels for each test set when running LP

[6] www.libama.org

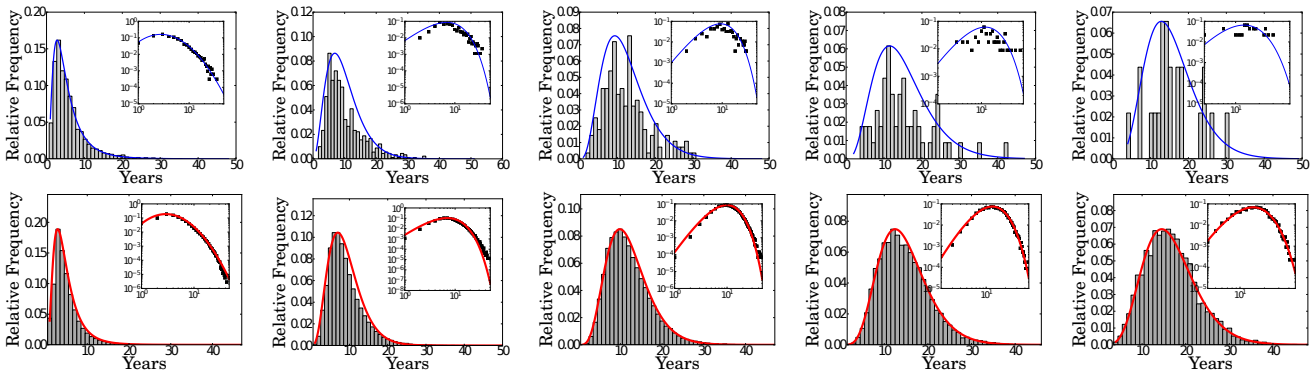


Figure 3: (top) Regularities in AAN. From left to right: The individual migration propensity for CS is log-normal. The kth move propensities ($k = 2, 3, 4, 5$) become gamma distributed for higher values of k. However, due to the small size of the AAN dataset, the data is sparse for higher values of k. (bottom) Regularities in DBLP. We find very similar patterns in the DBLP dataset with much more data available for all values of k. For the propensity ($k = 1$; most left panel) the log-normal has the best log-likelihood fit (significant at $p = 0.05$ according to a paired McNemar test on the continental as well as national levels). For $k > 1$ (other panels) the best fit in terms of log-likelihood is a gamma distribution, again significant (McNemar test, $p = 0.05$). (Best viewed in color)

for 200 iterations. Having access to only 36% of the labels and more, we can achieve an accuracy ≥ 0.80. As expected, reducing the number of labeled nodes slowly decreases the performance. With only 6% labeled nodes, we still achieve an accuracy of 0.58, which is a high performance on a multi-label problem with roughly 800 classes; the accuracy of random label would be 0.00125.

Next we investigated the reduction in run time due the compression of the LP graph on this dataset as well as the potential benefit of using a parallelized linear algebra library. Fig. 2(middle) shows the run times of LP and CLP for a varying number of cores. Here, the run time measures the total time required by the script. In the case of CLP, this includes the compression of the matrix as well. By using CLP instead LP, we reduce the run time by 20%. Of course, the relative run time is further reduced with more iterations, as the overhead of the compression vanishes. If we solely look at the run time of the power method, we see that CLP spends up to 25% less time on matrix-matrix-multiplications.

To investigate the scaling of the statistical regularities found, we considered **DBLP**. Compared to AAN, DBLP is roughly 100 times larger but contains relatively speaking way fewer labels. More precisely, the DBLP dump in use contains 1,894,758 papers written by 1,080,958 authors. However, of the 5,033,018 author-paper-nodes, only 10% are initially labeled with one of 4,350 cities. With a dataset of such size, running LP on a single machine is not easily possible anymore, even with modern hardware. With 4k+ labels, Y alone requires more than 160GB with 64bit float numbers. Additionally, the LAMA implementation did not work with arbitrary large matrices in our experiments. One way to overcome this barrier is to split the labels matrix into k chunks and do the multiplications separately. Afterwards, we can merge the results and obtain the final labeling. Using this approach, we compared CLP with LP to see how many cities each variant can handle in a single run. We started with the first 50 cities and added cities as possible. The results are depicted in Fig. 2(right) for 200 iterations of LP and show that CLP requires far less RAM because the matrices Q and J are much smaller than the uncompressed ones. We can now run LP with up to 570 cities in a single run and hence would only require 8 machines in a distributed setting. On the contrary, LP can only handle 425 at a time and would require 11 machines. The run times in Fig. 2(right) exclude the time needed for the compression which is negligible because we only have to compress T once and not for every chunk. The average total run time, including compression, for 200 iteration on the DBLP dataset with CLP is about 5.40 hours. This is a lot faster than LP which took on average around 7.49 hours. After running LP on both datasets, we are ready for an in-depth analysis of migration patterns in the augmented datasets.

5.2 Inferred Migration Regularities

The previous experiments have shown that CLP can help building large scale bibliographies augmented with geographic information. We will now use the two enriched datasets to infer statistical regularities within migration behavior of CS researchers. Unfortunately, we cannot directly observe the event of transfer from one residential location, respectively institution, to another one by a researcher. Instead, we use the affiliations mentioned in their publication record as a proxy. Nevertheless, even after running LP on the city level, this list may still be noisy and does not provide the timing information directly. To illustrate this, an author may very well move to a new affiliation and publish a paper with their old affiliation because the work was done while being with the old affiliation. Therefore, we are considering *migration sketches* as a proxy. Intuitively, a sketch captures the main stations in a researcher's career.

We define a migration sketch as the list of unique affiliations of an author ordered by the first appearance in the list of publications. This approach has the drawback that we cannot capture a researcher returning to an earlier affiliation after several years. Finally, we dropped implausible entries

from the resulting sketch database. For instance, we dropped sketches with more than ten affiliations because such sketches should be attributed to an insufficient entity disambiguation. Having the migration sketches at hand, we can now define *migration* or a *move* of a researcher as the event of transfer from one residential location to another one in the corresponding migration sketch. With these sketches, we can now start to investigate the statistical nature of researcher migration.

(R1) The propensity to transfer to a new research institution across scientists is log-normal: Let T_i^j be the point in time when researcher j makes their ith move from one location to the next one. Let t_i^j be the time between the T_{i-1}^j and T_i^j. T_0^j is the first temporal reference we have for an author, i.e., the year of their first publication listed in the bibliography. We call t_i^j, i.e., the time between two moves, the *(migration) propensity*. It reflects the bias of researchers to stay for a specific amount of time until moving on. Using maximum likelihood estimation for the parameters, we fitted the data to various distributions, including log-normal, gamma, inverse Gaussian, and Weibull. We found the log-normal distribution [Aitchison and Brown, 1957; Stewart, 1994] fitting the best in terms of log-likelihood; significant according to a McNemar test (p = 0.05). That is, the log of the propensity is a normal distribution with density $f(x) = (x\sqrt{2\pi\sigma^2})^{-1} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$. The parameters μ and $\sigma^2 > 0$ are the mean and the standard deviation of the variable's natural logarithm. This is a plausible model due to Gibrat's "law of proportionate effects" [Gibrat, 1930]: the underlying propensity to move is a multiplicative function of many independently distributed factors, such as motivation, open positions, short-term contracts, among others. Such factors do not add together but are multiplied together, as a weakness in any one factor reduces the effects of all the other factors. Moreover, a computer scientist stays on average 5.7 years at a place. Thus headhunters, should approach young potentials in their fifth year. One should maybe also reconsider the common practice of having projects lasting only three years. More importantly, the log-normality of the propensity can be found across continents and countries (results not shown here due to space restrictions) when we consider only moves originating from a continent, respectively country: there are no cultural boundaries.

(R2) k-th Move Propensities are Gamma: Fig. 3, columns 2-5, shows the best fitting distributions in terms of log-likelihood achieved by maximum likelihood estimation for the propensity to make $k > 1$ moves. More precisely, the kth move propensity for an author A_i is defined as $s_k^i = \sum_{j=1}^{k} t_j^i$. The best fit is a gamma distribution, $f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$ with shape $k > 0$, scale $\theta > 0$, and $\Gamma(k) = \int_0^\infty s^{k-1} e^{-s} ds$, suggesting that migration at later career stages is "memoryless"; again significant according to a McNemar test ($p = 0.05$). This conforms to the theory of Poisson processes, for which the inter-arrival times are independent and obey an exponential form. Consequently, the distribution of t conditioned on $\{t > s\}$ is again exponential. That is, the remaining time after we have not moved to a new position at time s has the same distribution as the original time t, i.e., it is memoryless. Moreover, we know that the time until the kth move — the kth move propensity — has a gamma distribution; it is the sum of the first k propensities of senior researchers. So, the propensities for the next move turn exponential for later career stages. Early career researchers have seldom taken many positions and we consider here rather senior researchers with typically permanent positions; they do not have to greatly care about their moves.

(R3) Brain Circulation is Gamma: Brain circulation, or more widely known as *brain drain*, is the term generically used to describe the mobility of high-level personnel. It is an emerging global phenomenon of significant proportion as it affects the socio-economic and -cultural progress of a society and a nation, and the world. Here, we define it as the time until a researcher returns to the country of their first publication. Only 29,398 out of 193,986 (15%) mobile researchers, i.e., researchers that have moved at least once, and out of all 1,080,958 (3%) researchers returned to their roots (in terms of publications) in the DBLP dataset. It also follows a gamma distribution. This indicates that returning is memoryless as well. Researchers cannot plan to return to their roots but pick up opportunities as they arrive.

6 Conclusion

International mobility among researchers not only benefits the individual development of scientists, but also creates opportunities for intellectually productive encounters, enriching science in its entirety, preparing it for the global scientific challenges lying ahead. Moreover, mobile scientists act as ambassadors for their home country and, after their return, also for their former host country, giving mobility a cultural-political dimension. So far, however, no statistical regularities have been established for the researcher migration within CS at global scale. One explanation might be that no transnational, registered dataset existed before.

We have demonstrated that harvesting and mining such a transnational, registered dataset from the Web is possible when using AI techniques such as label propagation (LP) to infer missing information. Such an enriched bibliography can be used to discover surprisingly simple and strong regularities. Actually, although not shown here due to space restrictions, there are no cultural boundaries underlying the timing events. The patterns remain similar no matter what region one looks at. Thus, moving on to a new position is a common pattern in terms of timing across different countries and independent of geography, ideology, politics or religion. Indeed, people have had the suspicion of many of these regularities but we have shown that they go beyond folklore. To find such regularities more quickly, we introduced a novel LP approach, called compressed LP (CLP), that runs LP on a compressed graph. We proved and demonstrated that CLP can significantly reduce the run time and memory consumption of LP without sacrificing performance at all.

Indeed, we have only started to look into migration through the "AI and the Web" lens. In the future, other AI techniques should be explored to reveal more migration patterns, and our results should be extended beyond CS. One should also investigate bootstrapping CLP via pseudo

evidence [Hadiji and Kersting, 2013] in order to reduce the required amount of memory for the labels matrix even further and in turn speed up convergence. In other LP tasks, lossless compression is unlikely to result in compression of the LP graph. Therefore, one should investigate lossy compressed LP as well.

Acknowledgments The authors would like to thank the anonymous reviewers for their feedback. The work was supported by the Fraunhofer ATTRACT fellowship STREAM and by the Deutsche Forschungsgemeinschaft (DFG), KE 1686/2-1, as part of the Coordination Project SPP 1527.

References

[Ahmadi *et al.*, 2011] B. Ahmadi, K. Kersting, and S. Sanner. Multi–evidence lifted message passing, with application to pagerank and the kalman filter. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.

[Aitchison and Brown, 1957] J. Aitchison and J. Brown. *The Lognormal Distribution*. Cambr. Univ. Press, 1957.

[Alexandrescu and Kirchhoff, 2007] A. Alexandrescu and K. Kirchhoff. Graph-based learning for phonetic classification. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 359–364, 2007.

[Bengio *et al.*, 2006] Y. Bengio, O. Delalleau, and N. Le Roux. Label propagation and quadratic criterion. In *Semi-Supervised Learning*, pages 193–216. MIT Press, 2006.

[Burke *et al.*, 2013] M. Burke, L.A. Adamic, and K. Marciniak. Families on facebook. In *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*, 2013.

[Cohen *et al.*, 2008] J.E. Cohen, M. Roig, D.C. Reuman, and C. GoGwilt. International migration beyond gravity: A statistical model for use in population projections. *PNAS*, 105(40):15269–15274, 2008.

[Fujiwara and Irie, 2014] Y. Fujiwara and G. Irie. Efficient label propagation. In *Proceedings of The 31st International Conference on Machine Learning (ICML)*, pages 784–792, 2014.

[Gibrat, 1930] R. Gibrat. Une loi des repartitions economiques: L'effet proportionelle. *Bulletin de Statistique General*, 19:469–514, 1930.

[Gonzales *et al.*, 2008] M.C. Gonzales, C.A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[Grohe *et al.*, 2014] M. Grohe, K. Kersting, M. Mladenov, and E. Selman. Dimension reduction via colour refinement. In *Proceedings of the 22nd Annual European Symposium (ESA)*, pages 505–516, 2014.

[Hadiji and Kersting, 2013] F. Hadiji and K. Kersting. Reduce and re-lift: Bootstrapped lifted likelihood maximization for map. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

[Katebi *et al.*, 2012] H. Katebi, K.A. Sakallah, and I.L. Markov. Graph symmetry detection and canonical labeling: Differences and synergies. In *Turing-100*, pages 181–195, 2012.

[Kersting, 2012] K. Kersting. Lifted probabilistic inference. In *Proceedings of 20th European Conference on Artificial Intelligence (ECAI)*, 2012.

[Mantegna and Stanley, 1995] R.N. Mantegna and H.E. Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376:46–49, 1995.

[Mladenov *et al.*, 2014] M. Mladenov, A. Globerson, and K. Kersting. Lifted message passing as reparametrization of graphical models. In *Proceedings of the 30th Int. Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2014.

[Neumann *et al.*, 2011] M. Neumann, B. Ahmadi, and K. Kersting. Markov logic sets: Towards lifted information retrieval using pagerank and label propagation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, 2011.

[Park *et al.*, 2013] J. Park, V. Barash, C. Fink, and M. Cha. Emoticon style: Interpreting differences in emoticons across cultures. In *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*, 2013.

[Radev *et al.*, 2009] D.R. Radev, P. Muthukrishnan, and V. Qazvinian. The ACL anthology network corpus. In *Proceedings of the ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, 2009.

[Rees *et al.*, 2009] P. Rees, J. Stillwell, P. Boden, and A. Dennett. A review of migration statistics literature. In *UK Statistics Authority, Migration Statistics: the Way Ahead, Report 4, Part 2*. UK Statistics Authority, 2009.

[State *et al.*, 2013] B. State, I. Weber, and E. Zagheni. Studying inter-national mobility through ip geolocation. In *Proceedings of the ACM Conference on Web Search and Data Mining (WSDM)*, pages 265–274, 2013.

[Stewart, 1994] J.A. Stewart. The poisson-lognormal model for bibliometriy/scientometric distribitions. *Information Processing and Management*, 30(2):239–251, 1994.

[Stillwell, 2005] J. Stillwell. Inter-regional migration modelling: A review and assessment. In *Proc. of the 45th Congress of the European Regional Science Association*, 2005.

[Zagheni and Weber, 2012] E. Zagheni and I. Weber. You are where you e-mail: using e-mail data to estimate international migration rates. In *ACM Conference on Web Science (WebSci)*, pages 348–351, 2012.

[Zhu *et al.*, 2003] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 912–919, 2003.

[Zipf, 1946] G.K. Zipf. The $p_1 p_2 / d$ hypothesis: On the intercity movement of persons. *American Sociological Review*, 11:677—686, 1946.