Gender homophily in online book networks

Doina Bucur

University of Twente, The Netherlands

Abstract

We measure the gender homophily (and other network statistics) on large-scale online book markets: `amazon.com` and `amazon.co.uk`, using datasets describing millions of books sold to readers. Large *book networks* are created by sales (two books are connected if many readers have bought both books) and can recommend new books to buy. The networks are analysed by the gender of their first author: is book consumption *assortative by gender*?

Book networks are indeed gender-assortative: readers globally prefer to read from one author gender (the global assortativity coefficients by gender is around 0.4). Although 33% of first authors among all books are female, female books are not proportionally sold together with male books: an average of 20% (and median of 11%) of books co-bought with male books are female books. Instead, female books make up on average more than half of the books co-bought with other female books.

The gender makeup of *literary genres* and structural *book communities* show that the gender homophily originates in a gender skew not only in certain literary genres (a fact known from prior studies), but even more strongly in certain book communities, with these book communities spanning multiple literary genres.

Keywords: Book, network, gender, homophily, genre, community

1. Introduction

Large, English-language online book markets sell individual books to individual readers, but also provide, on the webpages of most books, sales-based recommendations for other books. Two large groups of people (book authors, and their anonymous book readers) interact indirectly on these markets, via books. Book sales create salient ties among the books: if, over a period of time, a substantial number of the readers have bought both book *A* and book *B* on `amazon.com`, the website will record a sales-driven, co-buying tie between *A* and *B*. A link to book *B* will be posted by Amazon on the product page for book *A*, which will essentially act as a book recommender for future readers, and is likely to be self-reinforcing. The sales ties between books then form a *book network*: an information network which models our aggregated reading habits. Our global reading preferences can be studied over these large network models, more effectively than by surveying or measuring the reading habits of a small number of readers.

In this study, we focus on research questions related to gender: Do readers read books from both genders of authors (in other words, across all the readers' preferred literary genres, is there *diversity* between genders?) Is there instead a statistical preference of readers for one gender of authors, a preference which even spans literary genres, and which is measurable as *gender homophily* (or *gender assortativity*) in the book network? Are there gender-skewed clusters of books, or effective 'reading bubbles', for the readers on these online book markets?

These research questions are complex, and combine a number of simple questions, such as asking how many readers prefer exactly which subset of literary genres and in what proportion, and then asking which literary subgenres are gender-skewed in authorship. The related work has partial answers for these simpler questions: on a small number and size of datasets, prior studies have shown that gender preferences exist in what regards the writing of books (for example, there are literary genres where female authors dominate sales, and likewise for men), and the reading of books (male readers tend to appreciate certain literary subgenres, such as short stories or paranormal romance, at

Email address: `d.bucur@utwente.nl` (Doina Bucur)

different ratios than female readers). However, no prior studies exist which answer the question whether global gender homophily is present in the network of real-world book sales, and whether there exist global gender-skewed reading bubbles.

In content networks such as those formed by book sales, research questions related to gender are important. When clusters or communities of books in the book network are shown to be heavily segregated by the gender of their authors, then some readers are primarily exposed to authors of a single gender, thus removing the diversity of point of view which is otherwise available on the global book market[1]. We show here that strong gender homophily exists in book networks. Similar degrees of homophily were found before in other social networks [18], with the lower diversity in the opinion present in a group likely to reinforce itself in time and lead to a segregation of views [13]). We give detail to the related work in Section 2.

Summary of findings

We find that global gender assortativity in online book networks exists, and is partly 'baseline' homophily, namely caused by a difference in the ratios of male and female authors: female authors of English-language or English-translated books with high sales ranks on `amazon.com` and `amazon.co.uk` make up around 33% of all authors, and are thus less represented as authors than in the general population. Another 6% of authors are ungendered collectives.

Substantial homophily above the baseline is also present. The natural preference of readers for various subsets of *literary genres*, combined with the often skewed gender makeup of literary genres, and also combined with further reader preference within a genre, leads to a relatively high *global gender assortativity* (a nominal assortativity coefficient of 0.47-0.50, where a value of 0 is neutral, and a value of 1 is completely assortative). We also see substantial differences in the distributions of *local assortativity* metrics between male and female books (the mean and median ratio of ties to female books from male books are only around 20% and 11%, respectively, compared to 56-62% from female books).

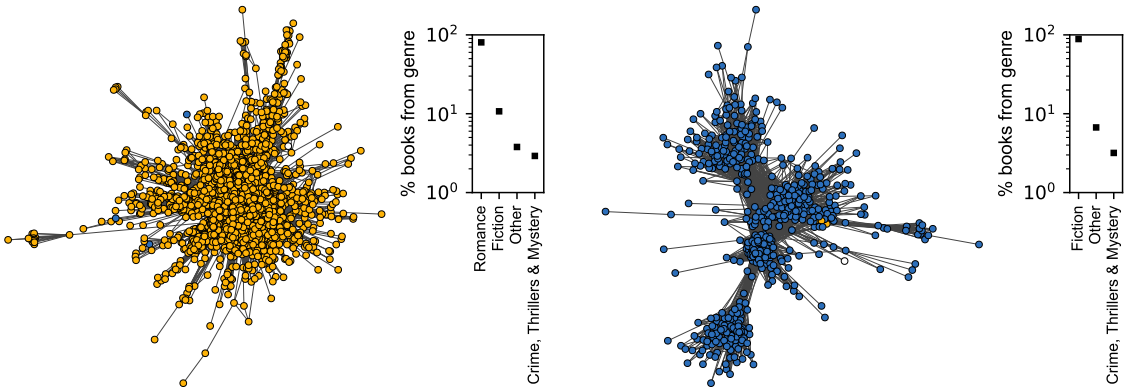

Figure 1: Female-authored books in orange, male books in blue, collectively authored books in white: examples of gender-segregated book communities based on `amazon.co.uk` book sales (British book communities C_3 and C_1; see Section 4 for more information)

Exclusively male and exclusively female *book communities* (so, reading bubbles) exist. These book communities are books clustered together solely based on the sales ties created by their readers (rather than clustered on other book information, such as its literary genre, or author name). Figure 1 shows two gender-segregated book communities among `amazon.co.uk` books: on the left, a community of 3328 fiction books (of which 99.85% female-authored), and their literary genres (of which 80% from the *Romance* genre); on the right, a community of 597 also fiction, but mostly male-authored books (0.50% female authors, with 89% of all books from the *Fiction* genre). A detailed analysis of the data is in Section 4.

[1]While our data does not reveal the gender of the book readers themselves, we conjecture, based on previous studies of book reviews [26], societal homophily [18] and cultural preferences [7], that the readers showing a very skewed preference for female-written books are themselves in most cases female, and likewise for men.

Data collection

We analyse book metadata (including a book's ISBN, author names, and literary genre) and the undirected book co-buying relationships for over 3 million print books. This study focuses on books on sale online by Amazon on the British and the American book markets—both large and mainly English-language markets, on which Amazon is a leading online seller. The data for a unique ISBN was crawled from the book's public Amazon web page. In many cases, the same book, with the same ISBN, is on sale on both national markets; separate crawls on `amazon.co.uk` and `amazon.com` then yield largely the same book metadata, but different co-buying relationships. In many cases, the same title is also sold as different editions, with different ISBNs, on the two markets, so not only the co-buying relationships will differ, but also some of the book metadata.

We collected British data (778 005 books) and American data (1 461 206 books) in the last quarter of 2017, specifically for this study. We denote these two datasets by $\mathbf{UK_{17}}$ and $\mathbf{US_{17}}$, respectively. A third, older, American-only dataset (filtered here down to 972 717 books) was acquired in an unspecified period between 1996 and 2014 for prior research [17], and is denoted here by $\mathbf{US_{14}}$.

$\mathbf{UK_{17}}$ and $\mathbf{US_{17}}$ were crawled starting from the top 100 best-sellers across all book genres, with the collection limited to books with good *sales ranks* (between 1 and 1 million, as reported by Amazon), plus all their immediate co-bought books, regardless of the sales rank of the latter. The relatively high sales rank, at least at one end of each tie between two books, ensures that the relationship between the books is significant, i.e., that a substantial number of readers have consumed both books. We cannot explain the difference in size between these two crawls. Since the internal structure of the networks is similar (for example, the average degree of the books in the two networks is similar), a likely explanation is the fact that, on the British Amazon, items other than print books have high sales rank, in a higher proportion than on the American Amazon.

The crawling methodology for the raw $\mathbf{US_{14}}$ dataset is not known; from the original dataset, we use all the books with at least one co-bought book (which is also present in the data), regardless of any sales ranks. Despite any differences in crawling method, we find consistent results across the three datasets.

Summary of method

We make use of established descriptive statistics from the area of network science [19], suitable for the empirical study of large network structures, including that of *information networks* (such as citation networks, the web, or *recommender networks* similar to our book networks) and online social networks. These descriptive statistics come in two categories, as follows. The simpler network metrics are statistics about individual vertices in the network, such as the vertex degree (number of co-bought books), or the local clustering coefficient, which describes the density of reciprocal connections in the immediate neighbourhood of a vertex. More complex network metrics describe instead the entire network. This is the case for the global coefficient of assortative mixing in the network (where the mixing is studied based on any 'category' of the vertex, such as its degree or, in particular, its gender). It is also the case for the global form of any metric describing individual vertices. For more insight on the choice of descriptive statistics for networks, see [19].

As a preliminary step, we run a *network analysis* to extract common network measures and metrics for these book networks, such as their degree distribution, their global and average local clustering metrics, and their global assortativity by vertex degree. The results, e.g., a non-assortativity by degree and a power-law distribution of vertex degrees, confirm prior measurements over other Amazon product networks.

Then, the gender study of the book networks starts with a process of *gender resolution*, in which 88% of the books in the datasets are assigned the true (rather than perceived) gender of their first author; the remaining books have gender unknown. The gender resolution is done using a number of corpora of first names, full names, and collective names (i.e., names of associations and institutions), all annotated with their gender, and reflecting the real population in the United States and the United Kingdom.

On the basis of gender-annotated vertices, we compute global and local *gender assortativity metrics*, and run statistical tests to show substantial differences in the local ties of, separately, male, female, and collective books. Finally, an analysis of the gender makeup of *literary genres*, and a structural analysis on the *communities* present in the book network (including their genre and gender makeup) show that gender homophily is local to certain genres and, in particular, it is local to certain book communities. A detailed description of this methodology is in Section 3.

2. Related work

In this section, we survey prior studies relevant to the main research question of measuring the homophily, or assortative mixing, in recommender book networks by the gender of the authors. Three categories of prior work are described below, and their current results are compared to those of this study.

Gender preferences when authoring, reading, and reviewing books

Gender in authoring. It is known that book authors do not have uniform success across the various literary genres: there are genres where female authors are more successful than male authors, and the reverse is also true. 2012 data on the volume of UK book sales of all time, containing only the sales numbers for the top 100 best sellers, shows that female authors sold more books than male authors in the general Fiction genre (and this genre overwhelms the list of best sellers) [8]. The top author, J. K. Rowling, is classified as writing in the subgenre Children's Fiction; the data shows this subgenre as being entirely female-written: all the bestselling authors in Children's Fiction are female. The opposite is true for another Fiction subgenre, Crime, Thriller & Adventure, where all the bestselling authors are men.

The gender of book authors was also surveyed more broadly, per literary genre, using a sample of 0.5 million English-language books crawled from `goodreads.com` (a website popular in the US) in [25]. Each author name was classified to a gender based on the first name, with ambiguous first-name instances removed from the data. This recent study shows that substantial differences in authorship exist in some genres, such as romance and comics books. Male authors are in the majority in most genres, except for children, adult, fantasy, suspense and cook books.

A similar case to book writing is the writing of pages in online encyclopedias, such as the free-content Wikipedia, where everyone can edit: a strong gender skew in the makeup of the editors of these pages is hypothesised to lead to systemic biases into the content of the encyclopedia. This has recently been confirmed: the narrow gender diversity of the Wikipedia editor community (under 10% of Wikipedia editors are women [10]) leads to notable women being covered and featured in many language editions of Wikipedia, but the way women are portrayed starkly differs from the way men are portrayed, with both structural and lexical gender bias present [28]. Women's pages are more linked to men's pages than vice versa. On a lexical level, romantic and family relationships are much more frequently covered in women's articles than in men's.

This study adds the following value to this existing work: our statistics strengthen the knowledge that certain genders of authors prefer (or simply are more successful in) certain literary genres, from the male-dominated Sports and Technology genres, to the female-dominated Romance. We show this using larger data than in any previous work, and in particular using data from different sources than before. As our Amazon data models all the real-world book sales, rather than the books popular on Goodreads, our statistics carry with a better degree of confidence.

Gender in reading and reviewing. For this study, we do not have access to data on the gender of the book readers themselves; only some prior work gained some knowledge on the question of how readers and writers associate in terms of gender.

The readers studied in [25] (namely, readers who wrote reviews on Goodreads) appear close to gender-balanced for most large literary genres. However, this result may not reflect the readers of books in the general population, since different ratios of male and female readers may choose to write public reviews, and also since Goodreads has a 76% female user base. Furthermore, genres overlap, are not well defined in all cases, and are composed of thematic subgenres whose readership may be gender-biased rather than balanced.

Also on a sample of 0.5 million Goodreads books, [26] measures the preference and appreciation of male and female readers towards (a) book genres, (b) the gender of authors, and (c) the book reviews written by fellow readers of one gender or another. The reviews available per book are ranked by Goodreads with an unknown algorithm, which may have lead to a sampling bias. Nevertheless, among the highly ranked reviews, male readers review more highly than females (and thus, may generally prefer) books from non-fiction genres, with the most male appreciation measured in genres such as biography, history, memoir and politics. Male readers also rate many fiction subgenres more highly than women: short stories, horror, crime, male-male romance and general literature. Female readers are instead seen to review more highly than men some romance subgenres (paranormal, contemporary, chick lit).

These results show that the literary genres have *internal 'gender lines'*, namely there exist subgenres strongly appreciated by only one gender (such as the female preference for paranormal romance), while the loosely defined genre as a whole may instead be more highly rated by male readers (a fact found to be true in this dataset for the

romance genre). Also in [26], Goodreads reviewers rate authors of their own gender more highly, in most book genres. This result is intuitive, and can be motivated by people's inherent preference for certain cultural themes which resonate best with their own life experience.

Gender homophily in book networks

While this is the first analysis of gender mixing in book networks, in our own preliminary work, we did a smaller-scale analysis of the oldest dataset also used here, $\mathbf{US_{14}}$, of around 1 million `amazon.com` books. Those preliminary results also showed strong local and global gender assortativity in book sales [6], and is included in this extended study.

Gender homophily in other information or social networks

In the general population, gender bias when forming ties, as well as a tendency to cluster by gender, have also been measured. For example, in an in-person Dublin community of hundreds of teenagers, males were predominantly in male-only clusters and females in female-only clusters, with male clusters usually larger [12]. In online social networks, the tendency towards gender segregation is clear, and more accentuated for male users [27] (on the gender-balanced Spanish social network Tuenti). In two samples of thousands of member profiles from MySpace, with a small female majority, both genders prefer as friends the majority gender, a choice more marked in females for their closest friend [24].

Gender homophily, both at and above the baseline, is documented, in various *professional* or other *content-based* social networks, in early studies (a number of which were surveyed in [18]): in a study (dated 1995) looking at political discussion networks, men have much higher levels of segregation than women, with 84% of men discussing politics only with other men. The authors of [18] conjecture that content-based relationships are more gendered than personal relationships, and work ties with men of status are used to gain advice, respect, mentoring, and access to information by both men and women. Other studies before 1997, surveyed in [18], also found that, at all levels in organizations, there are strong gender differences, with the minority gender having far more gender-balanced networks than the majority gender.

Newer studies (as recent as 2017) confirm the existence of skewed prosociality with gender in the networks formed by research professionals: [16] finds that, while researchers in general are prosocial, with 60%-80% sharing their own research material (publications and data), prosociality was most prominent from male to male, and less likely among all other combinations of genders. The authors conclude that this pattern suggests that male-exclusive networks exist in science, likely caused by an evolutionary history promoting strong male bonds. A study of collaboration structures in various engineering disciplines [11] also highlights a gendered scientific production, in which female engineers, although publishing in higher-impact journals, receive lower recognition (fewer citations) from the community. Both genders reproduce the male-dominated scientific structures by repeatedly collaborating predominantly with men. In collaboration structures in natural sciences, women are underrepresented in prestigious authorships compared to men, and this is accentuated in articles with the highest citation rates; there is a large negative correlation between the female representation in an authorship and the impact factor of the journal [2].

3. Method

The book network is an undirected graph $G = (V, E)$, where a vertex $v \in V$ is a print book, assigned an ISBN, and an edge (v, u) is an undirected also-bought relationship, i.e., either v was one of the books bought by the same Amazon customers as u (on Amazon's product web page for u), or vice versa. Either one or both Amazon web pages (for books v and u) may yield such an edge. Since Amazon limits the also-bought section of each web page to 100 products, both pages are valuable to retrieve these edges. There is no inherent directionality to this relationship, as the timeline of customers buying the books is not recorded on the web page of any book.

Of the metadata for each book, this study sets as vertex attributes the book's ISBN (e.g., *0099590085*, a best seller at the time of this study), the list of author names, if such a list is given on the product page (e.g., *Yuval Noah Harari*), and the book's category (or literary genre) as annotated by Amazon (e.g., the three-level *Science & Nature* ▹ *Biological Sciences* ▹ *Evolution*, from which we use the first level, here *Science & Nature*). We analysed literary

genres containing a substantial number of books (at least 500). The genre field is not present in the older $\mathbf{US_{14}}$ dataset, so genre-based analyses are only done over $\mathbf{UK_{17}}$ and $\mathbf{US_{17}}$.

We made an effort to acquire good-quality author records for the books in all three datasets. The book records in the $\mathbf{US_{14}}$ data did not include an author field at all; author records we crawled separately for this study from Amazon, by ISBN. The book records in the 2017 crawl ($\mathbf{UK_{17}}$, $\mathbf{US_{17}}$) had their author retrieved from Amazon. For all datasets, we collected a partial second set of author names by looking up each ISBN in the public OpenISBN records[2]. In a minority of cases, the Amazon first author was not identical with the OpenISBN first author; we carried forward the Amazon first author (as generally this was cleaner data), except in the cases when this was empty or equivalent (e.g., *Unknown*, *Not Available*).

3.1. Network analysis

We first analyse the three raw book networks structurally. We report standard graphs metrics such as the network size, the standard *global* [15] and average *local* [29] *clustering coefficients*, and Newman's *global assortativity coefficient* by vertex degree [21]. We also verify that the degree distribution fits to a power-law function[3].

The global clustering coefficient, in the range $[0, 1]$, is the probability that any two vertices are connected, conditional on them sharing at least one neighbour; a low value can indicate a network with relatively small clusters compared to the overall size of the network. The local variant of this coefficient computes this degree of transitive connection only within the neighbourhood of individual vertices; the average local clustering coefficient is higher than (and need not correlate with) the global version, as it weighs sparse neighbourhoods equally as dense ones (while in the global version, dense neighbourhoods weigh higher).

Newman's global assortativity coefficient is a standard metric which measures whether vertices in the graph prefer to link to (or mix with) similar, rather than dissimilar vertices, according to discrete characteristics of each vertex; in this case, the characteristic is the degree of each vertex. (Later, we also calculate this metric over another characteristic: the gender of book authors.) This global assortativity coefficient takes values in $[-1, 1]$; it is zero in a network where vertices link randomly, and takes a high positive value when high-degree vertices associate preferentially with other high-degree vertices (and likewise for low-degree vertices).

3.2. Gender resolution for the first author

The first step in the gender analysis resolves each first author into a discrete gender, i.e., assigns each vertex an attribute from the set $\{male, female, collective, unknown\}$. We aim for the true, rather then the perceived, gender of an author, so misleading pseudonyms (pen names), gender-neutral first names, and collective names formed based on the name of an individual are all assigned the true gender manually.

We first obtain a list of over 10 000 *collective* full author names present in any of the datasets; this set consists of authors such as *Oxford University Press*, *Cambridge International Examinations*, *Department of Agriculture*, *Editors of Bicycling Magazine*, or *Dorling Kindersley Publishing*, which cannot be assigned a definite gender. This list is extracted from the book datasets themselves, by filtering the author records for keywords which signal a collective (*Library*, *Press*, *Department*, *Editors*, etc.) and then manually inspecting the entries selected to remove names not belonging to this category (e.g., *Clare Press*, a female fashion journalist and author). Then, all books whose authors remain in this list are classified as collective. Doing this as a first step is crucial, because attempting to classify some of these collective names as if they were individuals (taking the first word as the first name) leads to incorrect gender resolution in cases such as *Peter Pauper Press*, *Herb Lester Associates Limited*, *Kenneth Grahame Society*, or *Marco Polo Travel Publishing*, which are publishers, rather than the male individuals they were named after. This is shown as **Step 1** in Figure 2.

The majority of the remaining authors are classified by extracting from the author's full name their first name, and querying gender-annotated corpora of first-name use in the real population (**Steps 3-4** in Figure 2). Three data sources are used:

[2] `www.openisbn.com`, accessed 2017.

[3] For power-law fitting, we use the Python `powerlaw` package, which implements the power-law fitting method from [9].

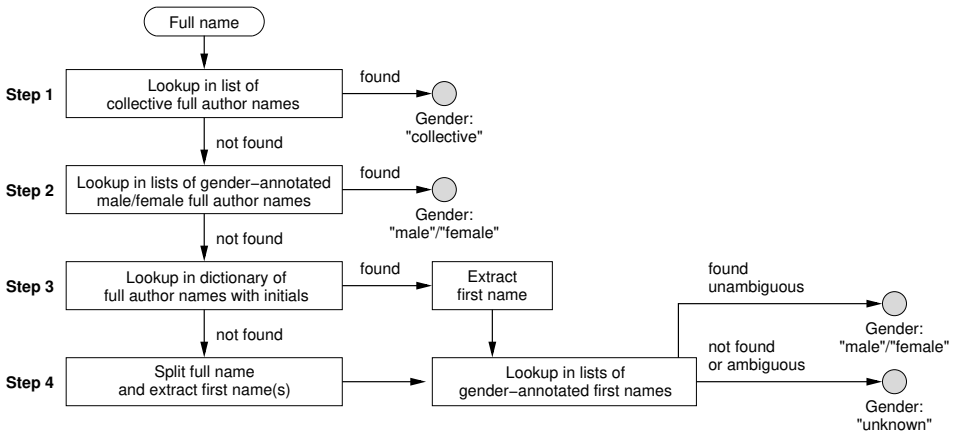

Figure 2: Steps in resolving the gender of an author

1. a corpus of annotated first names collected by the School of Computer Science at Carnegie Mellon University (CMU)[4], which categorizes approximately 5000 female and 3000 male first-name variations;

2. a probabilistic library[5] for gender detection based on first names, with data sourced across a number of years for all births in the United States and the United Kingdom;

3. our own list of manually gender-annotated first names from non-English-speaking countries, present among the authors whose books are sold (perhaps in translation) on the US and UK English-speaking markets (e.g., *Björn* is a decidedly male, and *Geneviève* a female first name).

This first-name lookup, if executed as such, will occasionally lead to incorrect classifications, as there exist cross-gender pen names (*James Tiptree Jr* is the pseudonym of the female science-fiction writer *Alice Bradley Sheldon*, and *Magnus Flyte* is the male pseudonym of a duo of female fiction authors), as well as ambiguous first names: over 300 first names from the CMU corpus are used by either gender (e.g., *Evelyn*, *Chris*, *Dana*). For this reason, **Step 2** in Figure 2 precedes the first-name lookup, and it categorizes the full names of those authors in the dataset who have cross-gender or ambiguous first names, by relying on the authors' surnames to make a difference.

For Step 2, we manually searched for Amazon or Internet author pages, or other concrete indication of the gender these authors subscribe to. These results make up new annotated lists of full author names (over 1000 female and 1500 male; e.g., *Dana Andrew Jennings* is a male American journalist and author, while *Dana Sachs* is a female American novelist). This list includes authors who use initials exclusively, but whose gender is known despite this (e.g., *P. K. Hallinan*, a male author of children's books). It also includes authors whose first names are nicknames, or are unusual and cannot otherwise be resolved to a gender, e.g., *Shoo Rayner* and *Crockett Johnson* are men's pen names, and *Yellow Tanabe* and *Banana Yoshimoto* are women's. The list also contains a small number of collective or anonymous authors whose gender is clearly predominantly female (e.g., *Girlguiding UK*) or male (e.g., *Men's Fitness*). Note that this gender-annotated author data may (in theory) have a small number of inaccurate entries, e.g., in the case when there exist two published authors with identical full names, an ambiguous first name, but different actual genders and books authored.

Other authors do use initials (at least occasionally), but their first names are publicly known (e.g., the art historian *E. H. Gombrich* is *Ernst Hans*). To be able to categorize the most popular such authors, we manually build a dictionary from author name with initials to complete author name, using Wikipedia and the wider Internet; this dictionary currently contains 400 names. This first name is then used for gender classification (**Step 3** in Figure 2). Two published authors with identical names using initials may exist, in which case we have aggregated both into the name of the most popular author.

[4] `www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/`, accessed 2017.
[5] `www.github.com/malev/gender-detector`, accessed 2017.

Finally, we evaluate the overall accuracy of this process of gender resolution in a manual verification step.

The outcome of the gender resolution is graph G with an added vertex attribute for the gender. We then denote by $G^k = (V^k, E^k)$ that subgraph of $G = (V, E)$ including only vertices of *known* gender, i.e., from which vertices of unknown gender (and any adjacent edges) have been eliminated ($V^k \subseteq V$ and $E^k \subseteq E$).

3.3. Global discrete assortativity by gender

Once the gender of the books is resolved and vertices have genders assigned, Newman's discrete assortativity coefficient r [21] by categorical vertex features is applicable to the book graph G^k.

The definition of r (equation (2) in [21], pasted as equation 1 below for completion) denotes by i or j any iterator over the vertex categories (over G^k, $i, j \in \{male, female, collective\}$). From among all edges in G^k, that fraction of edges which connect two books from any two gender categories i and j is denoted e_{ij}. The focus of the discrete assortativity coefficient is on assessing the actual occurrence of same-gender co-buying relationships when compared to random chance. For this, the term $\sum_i e_{ii}$ is the fraction of same-gender edges occurring in G^k. Then, a_i denotes that fraction of edges of which one end is a book from gender category i; for an undirected graph, this may be either end. In a randomly wired graph, the term $\sum_i a_i^2$ would give the fraction of occurring same-gender edges. r^k is the normalized difference between actual and random fractions of same-gender edges in G^k,

$$r^k = \frac{\sum_i e_{ii} - \sum_i a_i^2}{1 - \sum_i a_i^2}. \tag{1}$$

No assortative mixing is present in the graph if $r^k = 0$, such as the case of a randomly wired graph. If the graph is such that $\sum_i e_{ii} = 1$, then it has perfect assortativity, and $r^k = 1$. Perfect disassortative mixing occurs when $\sum_i e_{ii} = 0$ (and r^k is negative, but not necessarily -1).

An amount of bias may seep into the calculation of r^k if book buyers, after reading a book by author A, are more likely to consume books by *the same author A*, which raises the value of the term $\sum_i e_{ii}$. To remove that association bias, we also define r^d, a refinement of Newman's discrete assortativity coefficient: instead of over G^k, r^d is calculated using Equation 1 over the filtered graph $G^d = (V^k, E^d)$ from which same-author edges were removed, so any edge in E^d links two different author names ($E^d \subseteq E^k$).

3.4. Local discrete assortativity by gender

In the case of graphs where assortative mixing varies wildly from subgraph to subgraph, a local discrete assortativity coefficient by categorical vertex features is informative. For a vertex v in G^d, we denote v's one-hop edge neighbourhood (the set of all edges ending at v) by E^d_v, and v's one-hop edge neighbourhood of category i (the set of all edges between v and a vertex u of category i) by E^d_{vi}. Then, ϕ^d_{vi} is the fraction of edges local to v ending in a vertex of category i,

$$\phi^d_{vi} = \frac{|E^d_{vi}|}{|E^d_v|}. \tag{2}$$

We provide basic statistics for these local coefficients, specifically for the vertex category $i = female$. We use the simpler notation ϕ_v to mean $\phi^d_{v\,female}$. Table 1 summarizes the notation presented in this section.

Table 1: Notation

$G = (V, E)$	full book network
$G^k = (V^k, E^k)$	G including only books with known gender, $V^k \subseteq V$, $E^k \subseteq E$
$G^d = (V^k, E^d)$	G^k excluding same-author edges, $E^d \subseteq E^k$
r^k	global discrete (gender) assortativity coefficient over G^k
r^d	global discrete (gender) assortativity coefficient over G^d
ϕ_v	local discrete (*female*) assortativity coefficient for vertex v in G^d

We compare the three samples for ϕ_v resulted from segmenting V^k into three disjoint subsets, by gender (so, v will denote in turn a male, female, or collective book). A Kolmogorov-Smirnov two-sample test [30] computes a numerical distance D between the empirical distribution functions of any two of these samples, with the null hypothesis that the

samples originate from *the same distribution*. The test is non-parametric, and distribution-free. We run the test over all three pairs of ϕ_v samples, male vs. female, collective vs. female, and collective vs. male books. If the null hypothesis is rejected by the test (low p value and a relatively high D statistic above 0), books from different genders associate differently in terms of local ties to (here) female books.

3.5. Book community detection by co-buying relationships

To identify the structural communities in our large book networks, we use the efficient multilevel (Louvain) community-detection algorithm [3]. Maximizing the modularity of a partitioning (a value between -1 and 1 measuring the density of links inside as compared to between communities) is computationally hard [4]. Due to this fact, and the relatively large size of our book networks, a desirable performance factor for an algorithm for community detection is its complexity; the Louvain algorithm was shown to be particularly efficient over multi-million-vertex networks [3].

The Louvain technique is an iterative and hierarchical approximation suitable for multi-million vertex graphs: it first computes small communities by optimizing the local modularity in the neighborhood of each node; these are then modelled as 'supernodes', with the original graph becoming a smaller, weighted graph, which is also partitioned. The process repeats until the modularity of the partitioning doesn't increase.

We report the modularity of the partitioning, and analyse the book communities obtained in terms of size and size distribution, gender makeup and literary genre(s). The set of literary genres are largely, but not entirely, identical between the US and the UK markets. While an intuitive assumption is that book communities map well over literary genres, in our preliminary study [6] we observed that this was not necessarily the case: book communities are multi-genre, although in some cases one genre dominates. Due to this, the mapping between communities and literary genres is two-dimensional.

4. Results

4.1. Network analysis: clustering, degree distribution, and degree assortativity

Table 2 summarizes basic network statistics for the three datasets; all metrics pertain to the raw book networks (three instances of the undirected graph G), before resolving the gender of the first author. All networks are mostly connected. Despite the differences in size and crawling methodology among the book networks, their average degrees and clustering coefficients are consistent, and match the four generic Amazon co-purchasing product networks publicly available at the time of writing in the Stanford Large Network Dataset Collection [14].

Table 2: Network statistics (over graph G)

	$\mathbf{UK_{17}}$	$\mathbf{US_{17}}$	$\mathbf{US_{14}}$
network size $\|V\|$ (number of books)	778 005	1 461 206	972 717
number of co-buying relationships $\|E\|$	16 182 063	32 347 573	17 588 632
maximum degree of a book	4 618	8 323	5 739
degree average, stdev	41.59 ± 73.87	44.28 ± 89.86	36.16 ± 68.38
size of largest connected component	777 792	1 461 081	931 318
global clustering coefficient	0.151	0.130	0.158
average local clustering coefficient	0.433	0.439	0.429
global assortativity coefficient by degree	-0.037	-0.036	-0.031

The clustering coefficients (low in the global, and medium in the local variant) indicate more connectivity in the neighbourhoods of high-degree books, so the presence of vertex clusters in the network, with relatively small clusters compared to the overall network size. (An analysis of the concrete graph clustering is shown later in Section 4.4.)

Newman's degree-based assortativity coefficients (with values close to -0.03) show mostly non-assortative mixing by degree. This is akin to other real-world networks, such as the web (0.065 computed over a graph of 269 504 undirected hyperlinks among web pages from a single domain, in [20]), but is unlike the positive, and sometimes high (up to 0.400) assortativity measured in almost all human social networks, which is likely due to a higher transitivity of connection in human societies [22].

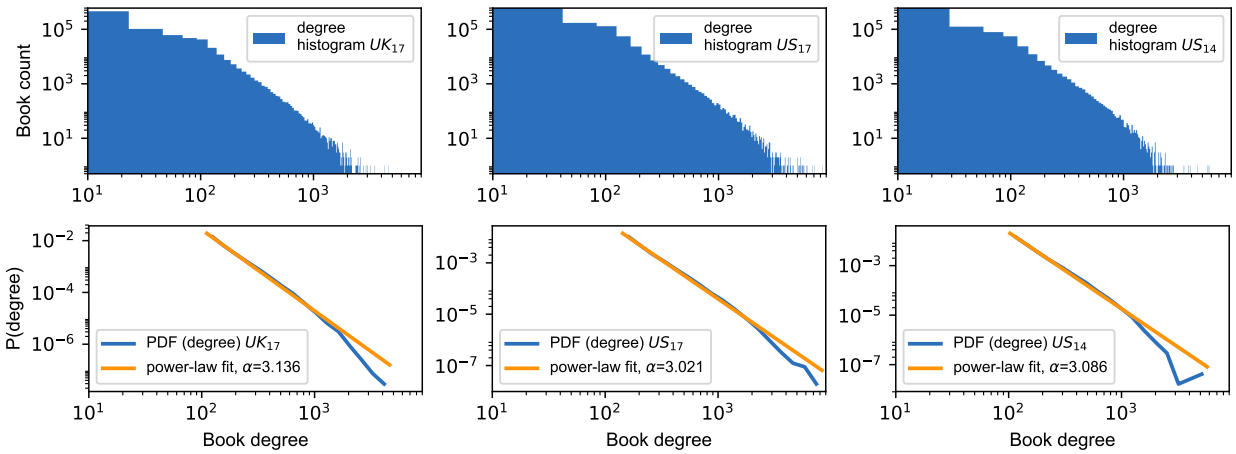

Figure 3: Degree histograms, probability density functions (PDF), and power-law fit

The three degree distributions on the raw books networks fit to power-law distributions, i.e., their probability density function (PDF, or normalised histogram) can be described asymptotically by $P(d) \sim d^{-\alpha}$, meaning that the likelihood that a vertex has degree d in this network decreases by a power of d. Figure 3 shows the three degree distributions as histograms (top row) and also as a PDF and its power-law fit, with exponents α around 3.1 and a good fit over the empirical data for mid-range degrees above 100 (bottom row). This holds for other real-world networks, such as the web, with most exponents between 2 and 3 [1]. Thus, online book networks are scale-free: they lack a characteristic degree, as the vertices have dominating low degrees, but also a long tail of high degrees.

4.2. Gender analysis: statistics and assortativity

Around 88% of the books had their gender resolved (Table 3); the remaining have their gender annotated as *unknown* in graph G, and are removed to construct G^k and G^d.

Table 3: Gender statistics (over graphs G^k and G^d)

	UK_{17}	**US_{17}**	**US_{14}**
% books with known gender	88.2%	87.9%	88.0%
number of books with known gender $\lvert V^k \rvert$	686 507	1 284 512	855 519
number of co-buying relationships $\lvert E^k \rvert$	13 125 164	26 507 425	14 421 568
gender ratios in G^k (% of male, female, collective books)	61.6% 32.4% 6.0%	60.7% 33.2% 6.1%	61.0% 33.8% 5.2%
global gender assortativity r^k	0.471	0.498	0.485
number of co-buying relationships $\lvert E^d \rvert$	11 233 395	22 278 441	12 636 474
global gender assortativity r^d	0.370	0.392	0.405
average local gender assortativity ϕ_v (among male, female, collective books)	0.198 0.562 0.234	0.203 0.577 0.250	0.197 0.590 0.257
median local gender assortativity ϕ_v (among male, female, collective books)	0.111 0.571 0.143	0.115 0.600 0.149	0.103 0.619 0.163

We evaluate to what extent the gender resolution is correct, over random samples of 100 books (whose first authors were resolved into a gender), one from each of the three datasets. We verify the gender results for these records manually; of the 300 instances, we found 2 wrong calls: a French first name whose gender is different in English than in French, and a collective name misinterpreted as an individual person. We conclude that, while incorrect instances exist, their numbers are not significant, and do not threaten the validity of the results.

Among the books with known gender, a majority (over 60%) have a male first author, and around 33% a female first author (Table 3). Newman's global coefficients of assortativity by gender (r^k and r^d, i.e., with and without same-

author links) are above 0.47 and 0.37, respectively: the book networks display overall assortativity by gender, i.e., preferential attachment of same-gender books, above random chance.

The local coefficients of assortativity with *female* books (ϕ_v) in G^d vary substantially with the gender of vertex v itself. On average, only roughly 20% of the books co-bought with books v written by male first authors have female first authors; the median is lower, at 10-11% across the three datasets. In contrast, both the average and the median ϕ_v over books v by female authors are 56-62% (Table 3, where the ϕ_v values are given as ratios in [0, 1]).

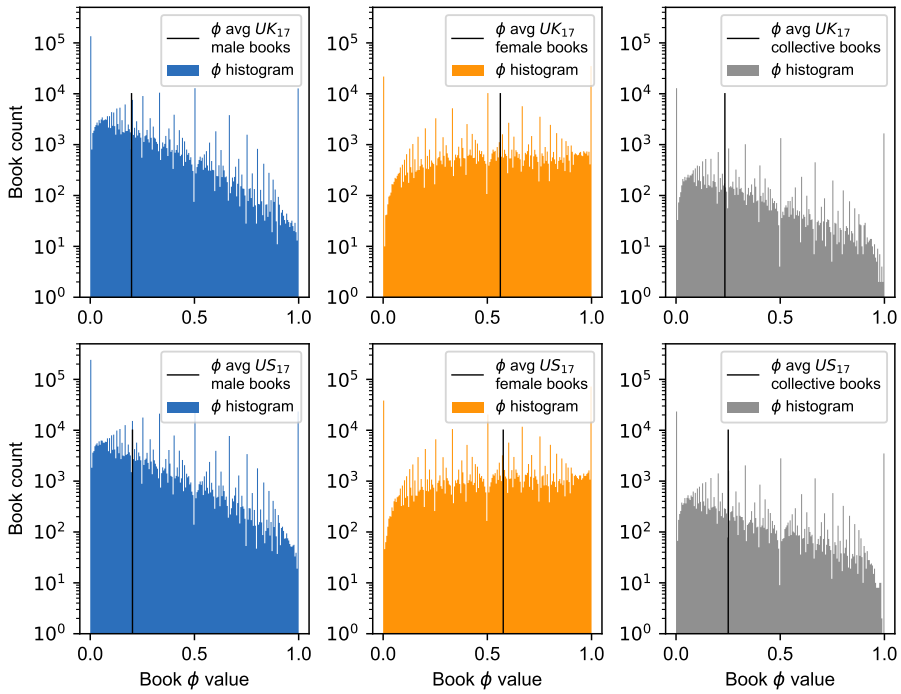

Figure 4: Histograms of three ϕ_v samples: books v of male (left), female (center) and collective (right) gender. $\mathbf{US_{14}}$ is similar and omitted.

The histograms of the three ϕ_v samples are shown in Figure 4; $\mathbf{US_{14}}$ is similar, and is omitted from the figure. The distribution of ϕ_v where the gender of v is male (Figure 4, left) is in contrast with that where the gender of v is female (center) and collective (right). The Kolmogorov-Smirnov two-sample tests (statistics in Table 4) reject the null hypothesis that male and female (also: collective and female) books associate similarly, with a D statistic above 0.5 in all cases.

Table 4: Kolmogorov-Smirnov statistics: distance D (with confidence level p) between the empirical distributions of ϕ_v where v are male, female, or collective books

	$\mathbf{UK_{17}}$	$\mathbf{US_{17}}$	$\mathbf{US_{14}}$
distance male-female	$D = 0.500, p < 0.001$	$D = 0.506, p < 0.001$	$D = 0.517, p < 0.001$
distance collective-female	$D = 0.445, p < 0.001$	$D = 0.434, p < 0.001$	$D = 0.429, p < 0.001$
distance male-collective	$D = 0.059, p < 0.001$	$D = 0.074, p < 0.001$	$D = 0.094, p < 0.001$

4.3. Genre analysis: gender distribution with literary genre

Among the books with resolved gender (graph G^k), in both $\mathbf{UK_{17}}$ and $\mathbf{US_{17}}$, there are a total of 34 literary genres with at least 500 books per genre. We aggregate all books in G^k either (a) without a stated genre, or (b) in minor genres containing under 500 books into the default genre called *Other*. Amazon's American genres differ slightly in name from the British ones (e.g., the British *Crime, Thrillers & Mystery* aligns to the American *Mystery, Thriller & Suspense*).

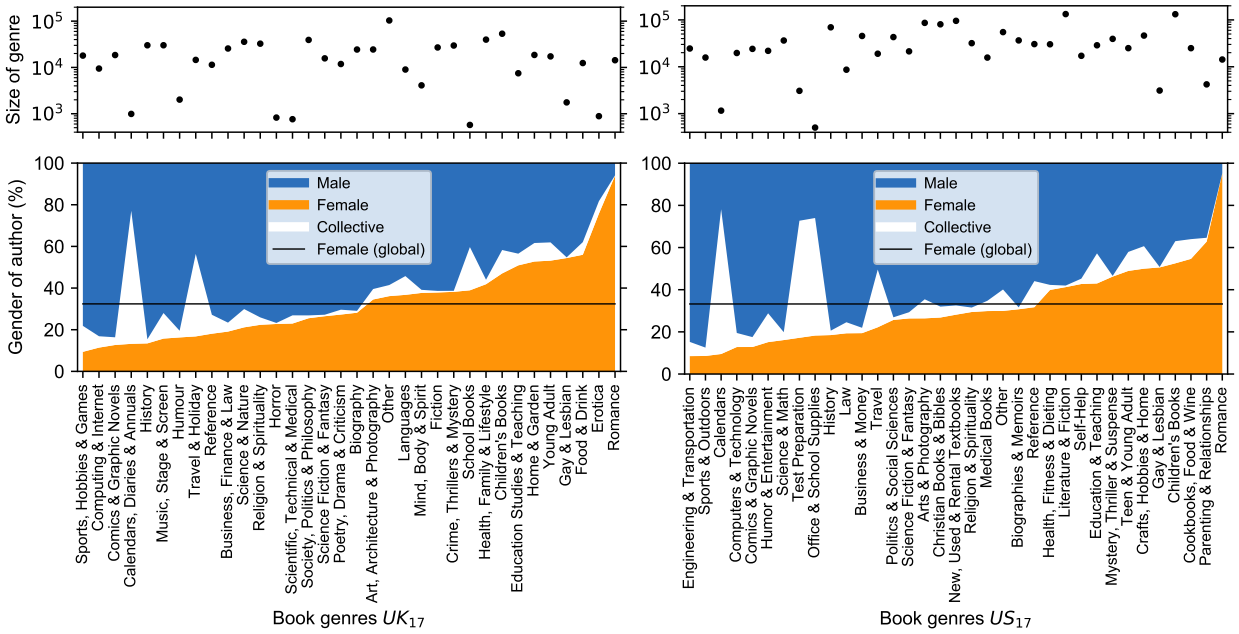

Figure 5: The size and gender breakdown of major literary genres in $\mathbf{UK_{17}}$ and $\mathbf{US_{17}}$

Figure 5 (bottom) shows the breakdown of each literary genre by the gender of the first author, for both $\mathbf{UK_{17}}$ and $\mathbf{US_{17}}$; the genres are shown in ascending order of the percentage of female first authors per genre, and in comparison to the global ratio of female authors across all genres in G^k (shown as a horizontal line). Figure 5 (top) adds the size of each genre, i.e., its book count; these counts add up to the size of graph G^k. A small minority of relationship-oriented genres are largely female: British *Romance* (approximately 14 000 books) is 94% female, and American *Romance* (also 14 000 books) is approximately 95% female. On the other hand, British *History* (30 000 books), *Comics & Graphic Novels* (19 000 books), *Computing & Internet* (9 000 books), and *Humour* (2 000 books) are all 80-85% male. Similarly, American *Sports & Outdoors* is 87% male, with *Computers & Technology*, *Comics & Graphic Novels*, *Science & Math*, and *Engineering & Transportation* also above 80% male.

4.4. Community analysis: preferential gender and genre in book communities

While some of the literary genres are unsurprisingly gender-skewed in authorship, a fact also found earlier in [25], this only translates into "bubbles" of authorship from a single gender only if readers effectively consume books limited to one or a few genres. Our analysis of book communities over graph G^k finds that while (a) genres are not isolated in the readers' buying preferences (i.e., different genres do get co-bought by readers), (b) the assortativity by gender found in Section 4.2 is due to the existence of cross-genre book communities, with some of these book communities even more gender-skewed than individual genres.

For the community analysis, the Louvain community-detection algorithm partitions the graphs G^k, with high values for the modularity metric: 0.86, 0.83, 0.84 for $\mathbf{UK_{17}}$, $\mathbf{US_{17}}$, and $\mathbf{US_{14}}$, respectively. This process of community detection comes with limitations: (a) the problem of extracting communities from such large graphs likely has multiple similar solutions, all of very similar modularity coefficients, and (b) the algorithm is a greedy heuristic, so does not necessarily produce an optimal partition of the graph. In this respect, we experimented with running the community detection repeatedly over the book networks; the resulting solutions occasionally merged a few of the clusters, but always obtained nearly indistinguishable, high modularity values, with insignificant change in the most gender-segregated communities.

Figure 6 shows the sizes of all communities found (including those of small size); the number of communities (data points Figure 6) is low, leading to data sparseness, so the histogram of community sizes (Figure 6, top), is

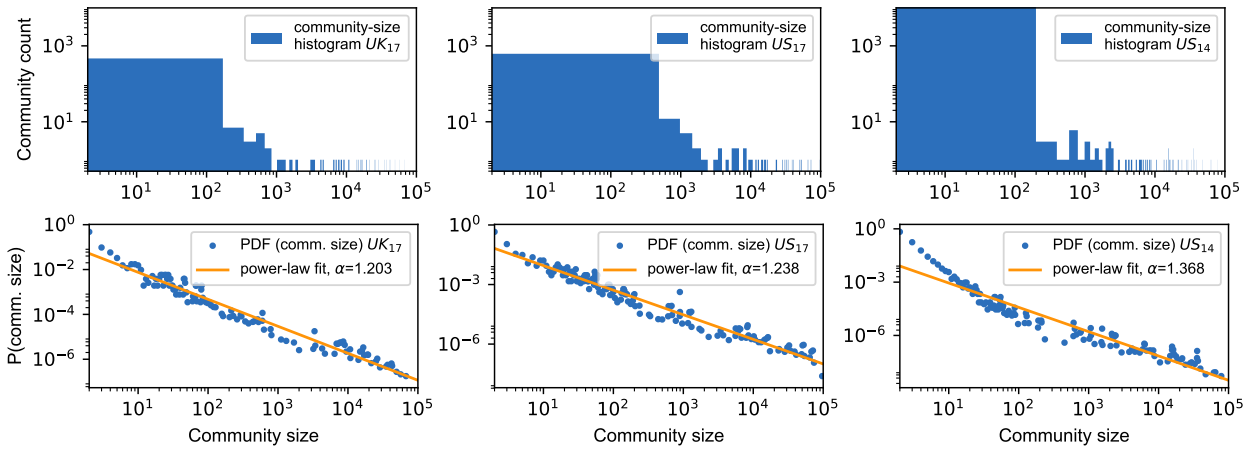

Figure 6: Community-size histograms, probability density functions (PDF), and power-law fit

first smoothed into a probability density function (PDF)[6], which is then fitted to a power-law density function with exponents α slightly above 1 (Figure 6, bottom). Thus, asymptotically, communities (like vertex degrees) are also scale-free and lack a characteristic size.

We study all communities larger than 500 books. This amounts to 45 communities for $\mathbf{UK_{17}}$, 61 for $\mathbf{US_{17}}$, and 57 for $\mathbf{US_{14}}$. Maps from community to literary genre (for $\mathbf{UK_{17}}$ and $\mathbf{US_{17}}$) are in Figure 7; this shows which ratio of books from each book community comes from which major genre.

Some book communities are dominated by one literary genre; e.g., the British nonfiction book community C_8 (2000 books) consists mostly of books from the *Sports, Hobbies & Games* category (92%). Other book communities are spread across genres; e.g., in the British book community C_{25} (4000 books), several nonfiction literary genres are represented substantially, e.g., *Languages* (43%), *Society, Politics & Philosophy* (25%), and *Reference* books (9%). These two communities are visualised in terms of gender and genre makeup in Figure 8. Two other examples (the British fiction-book communities C_1 and C_3) were given in Figure 1 (of Section 1).

Finally, the gender makeup of all book communities is shown in Figure 9, for the three datasets. In comparison with the gender makeup of literary genres (Figure 5), notable is the fact that there exist book communities which are more gender-segregated than all literary genres. The lowest percentage of female authors in a British literary genre is 9.38% (in *Sports, Hobbies & Games*), yet the lowest percentage of female authors in a British book community is 0.50% (in C_1, shown in Figure 1, right). Compared to the most female British literary genre (*Romance*, 93.95% female), British C_3 (Figure 1, left) is 99.85% female. Similarly for male authors: while the most male British genre is *History*, at 84.60% male, the most male British book community is C_1, at 99.33% male.

The graph data and further visualisations for this study are publicly available [5].

5. Discussion and conclusions

This study measured, using both global and local graph metrics over large graphs of online book co-sales, the degree to which book readers will read the work of authors from both genders. There are limitations to the methodology: the process of resolving the gender of the main author in millions of book records is inherently imperfect (e.g., non-English author names and names with initials are particularly likely to remain with unresolved gender, and a small fraction of wrong calls in gender resolution exist).

We found gender homophily (or assortative mixing) above the expected baseline given by the overall lower ratio of female authors. In particular, the average reader of female-authored books prefers to buy many other female books

[6]The smoothed PDF of community sizes is computed as follows. Any data point (x, y) from the histogram of community sizes is normalized into a PDF data point. Then, if this data point is adjacent on the x axis to intervals without data points, it is weighted down in y value proportionally to the width of the interval without data.

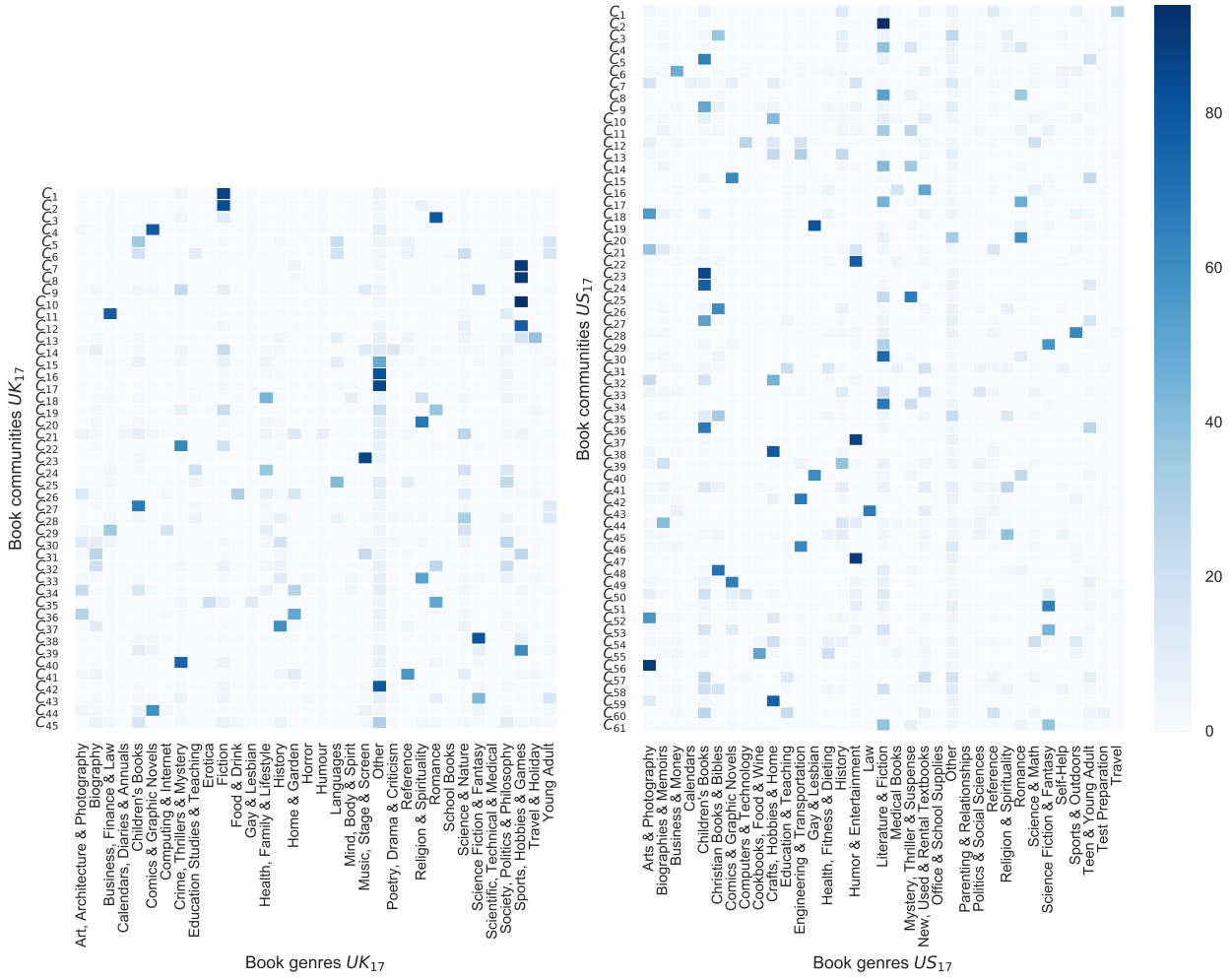

Figure 7: Which percentage of a community comes from which literary genre? Maps from communities to literary genres

– over half of the total number of books co-bought, a fraction which is remarkably close to the ratio of females in the general society, and which essentially corrects the effect of the baseline homophily. On the other hand, the average reader of male books prefers to buy substantially fewer female books than the baseline, thus reinforcing the gender homophily.

We measured the gender makeup of all major literary genres, and (expectedly) found the sports, technical, and comic-book categories to be substantially male, and the relationship-related, children, and cook-book genres substantially female. Less expectedly, we found structural, sales-driven book communities which are even more binary than individual genres: some communities are made entirely by one author gender, with examples of extreme local polarization within the main fiction literary genres.

These numerical results signal the existence of a substantial number of texts which are written for, marketed to, recommended for, or simply appeal to, a single gender of readers – likely the same gender as that of the authors of these texts. This supports the idea that the culturally established "binaries" of gender [23] in previous generations are still present, and reflect into book consumption. We leave for future work the analysis of which type of content, within a literary genre, correlates with a gender-skewed readership.

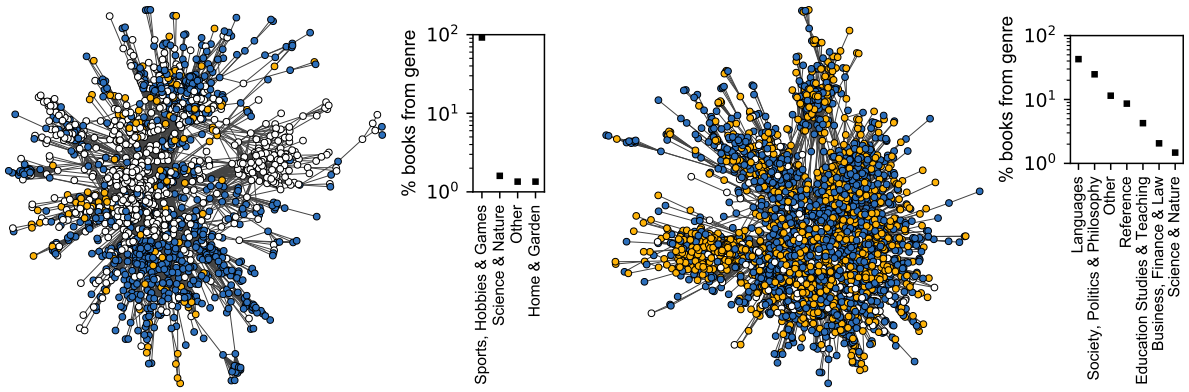

Figure 8: Female-authored books in orange, male books in blue, collectively authored books in white: examples of book communities based on `amazon.co.uk` book sales; the largely single-genre community C_8 (left) and the multi-genre community C_{25} (right)

References

[1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[2] Michael H. K. Bendels, Ruth Müller, Doerthe Brueggmann, and David A. Groneberg. Gender disparities in high-quality research revealed by Nature Index journals. *PLOS ONE*, 13(1):1–21, 01 2018.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. Maximizing modularity is hard. *arXiv preprint physics/0608255*, 2006.

[5] Doina Bucur. Network datasets. `www.cs.utwente.nl/~bucurd/data/`, February 2018.

[6] Doina Bucur. On the gender of books: Author gender mixing in book communities. In Chantal Cherifi, Hocine Cherifi, Márton Karsai, and Mirco Musolesi, editors, *Complex Networks & Their Applications VI*, pages 797–808, Cham, 2018. Springer International Publishing.

[7] Georgia T Chao and Henry Moon. The cultural mosaic: A metatheory for understanding the complexity of culture. *Journal of Applied Psychology*, 90(6):1128, 2005.

[8] Lynn Cherny. UK bestsellers: Remash by genre and gender, 2012. Accessed Oct 2018.

[9] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[10] Heather Ford and Judy Wajcman. 'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap. *Social studies of science*, 47(4):511–527, 2017.

[11] Gita Ghiasi, Vincent Larivire, and Cassidy R. Sugimoto. On the compliance of women engineers with a gendered scientific system. *PLOS ONE*, 10(12):1–19, 12 2016.

[12] Deirdre M Kirke. Gender clustering in friendship networks: some sociological implications. *Methodological Innovations Online*, 4(1):23–36, 2009.

[13] Gueorgi Kossinets and DuncanJ. Watts. Origins of homophily in an evolving social network. *American Journal of Sociology*, 115(2):405–450, 2009.

[14] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[15] R. Duncan Luce and Albert D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, Jun 1949.

[16] Jorg J. M. Massen, Lisa Bauer, Benjamin Spurny, Thomas Bugnyar, and Mariska E. Kret. Sharing of science is most likely among male scientists. *Scientific Reports*, 7(12927), 2017.

[17] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA, 2015. ACM.

[18] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[19] Mark Newman. *Networks*. Oxford University Press, 2018.

[20] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[21] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[22] Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003.

[23] Rebecca Solnit. *The mother of all questions*. Haymarket Books, 2017. Section *80 Books No Woman Should Read*, based on the 2015 essay `https://lithub.com/80-books-no-woman-should-read/` (accessed Feb 2018).

[24] Mike Thelwall. Social networks, gender, and friending: An analysis of MySpace member profiles. *Journal of the Association for Information Science and Technology*, 59(8):1321–1330, 2008.

[25] Mike Thelwall. Book genre and author gender: Romance paranormal-romance to autobiography memoir. *Journal of the Association for Information Science and Technology*, 68(5):1212–1223, 2017.

[26] Mike Thelwall. Reader and author gender and genre in goodreads. *Journal of Librarianship and Information Science*, 0(0):0961000617709061, 2017.

[27] Yana Volkovich, David Laniado, Karolin E Kappler, and Andreas Kaltenbrunner. Gender patterns in a large online social network. In *International Conference on Social Informatics*, pages 139–150. Springer, 2014.

[28] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a Man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *ICWSM*, pages 454–463, 2015.

[29] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440, 1998.

[30] I T Young. Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. *Journal of Histochemistry & Cytochemistry*, 25(7):935–941, 1977. PMID: 894009.

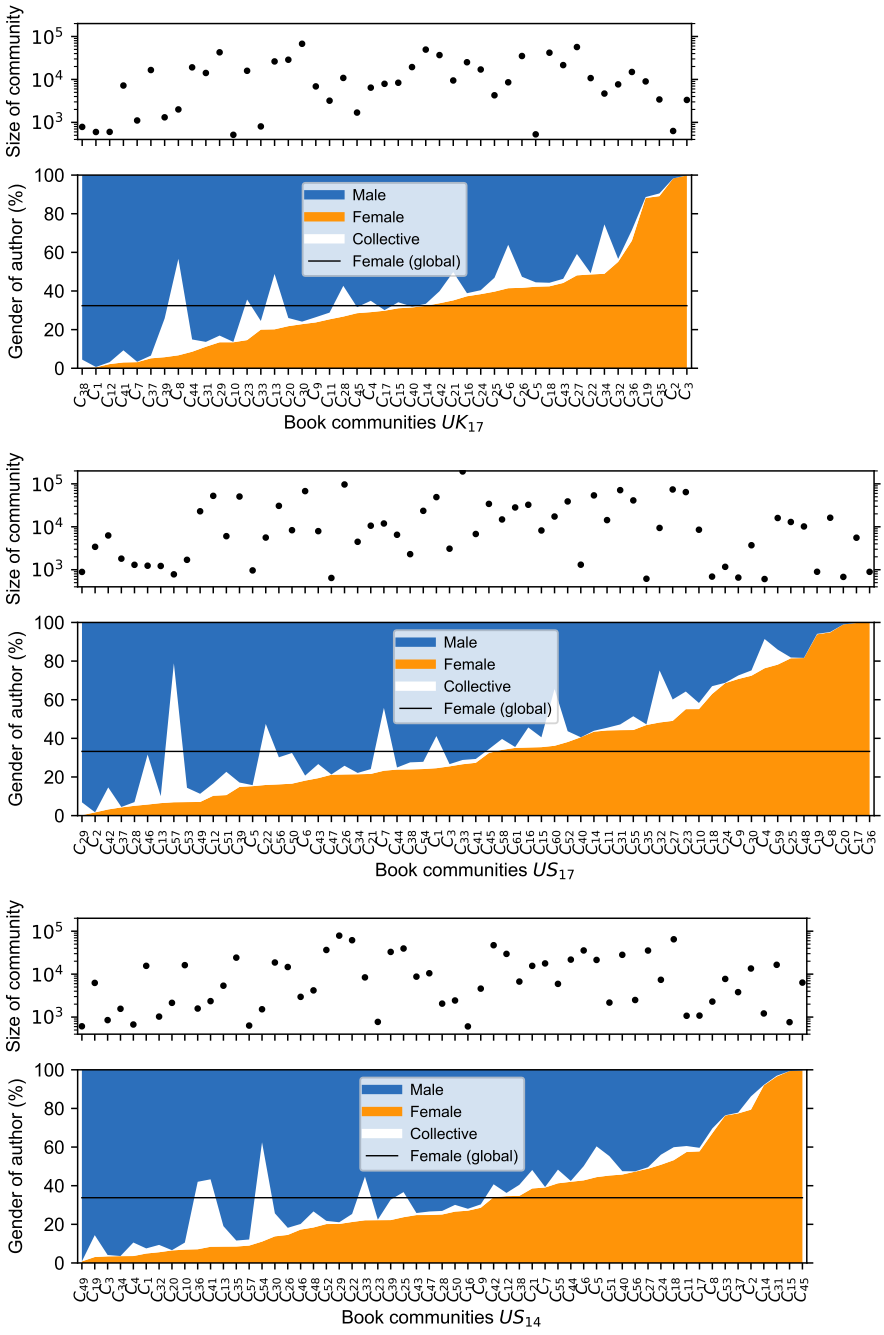

Figure 9: The size and gender breakdown of book communities in $\mathbf{UK_{17}}$, $\mathbf{US_{17}}$, and $\mathbf{US_{14}}$