On the gender of books: author gender mixing in book communities

Doina Bucur

Abstract Using a book co-buying network from amazon.com of over 1 million books, we find empirically that readers who have purchased male first authors before are substantially less likely than expected to buy books by female first authors, when aggregated across the entire book market. Conversely, past buyers of female authors are slightly more likely than expected to buy other female authors. This same-gender assortativity is found to be local: certain writing genres are "coloured" preferentially by one gender. This can be attributed both to writer availability (i.e., a gender's preferential attachment to writing for one genre), and to the buyers' preferential attachment to the output of writers of one gender. We obtain these insights by classifying the gender of the first author for most of the books, then running statistical tests which compare the gender makeup of books co-bought with either male or female books. Structural book communities, as generated from readers' co-buying choices, are computed, visualised in terms of gender makeup, and their writing genres are summarised to match the genre with a gender makeup.

1 Introduction

The commercial success of a writer lies with both their book's topic, and with the buying public. Certain genres (e.g., cooking books) may be dominated by writers of one gender due to these writers' own proficiency with the genre. Readers of a book genre may add to this gender domination by stereotyping the genre, and thus strongly favouring a certain author gender, then reflecting this bias in their buying habits. To overcome the gender stereotyping introduced by the readers, writers of either gender have resorted to using initials (*J. K. Rowling* is Joanne Rowling) or pseudonyms (*Robert Galbraith*, also for Joanne Rowling) in order for their books to appeal to a certain client demographic on equal footing.

Doina Bucur
University of Twente, The Netherlands. e-mail: d.bucur@utwente.nl

Using a large dataset of over 1 million books bought by hundreds of millions of readers in the recent history of amazon.com purchasing, together with the books' co-buying relationships as determined by the readers, we investigate whether de facto book genres nowadays show a balanced *gender mixing*. We answer in the negative: readers who already read male authors are substantially less likely to read female authors than expected. There exist book communities strongly dominated by male authors, as well as a few exclusive female communities, i.e., a local *same-gender preferential association* of the genders in certain writing genres. Whether due to a writer- or reader-side bias, this preferential gender association in book consumption echoes recent findings in other areas: academic prosociality was found to be most prominent from male to male researchers [8], and male-dominated scientific structures in engineering are formed by male scientists collaborating predominantly with men [3].

Dataset. We analysed metadata for 1,748,925 ISBN-assigned books from the largest online book seller, amazon.com, containing product recommendations (i.e., customer-mined co-buying relationships indicating the cascading of commercial writing success across books). The dataset was crawled from public Amazon webpages between 2009 and 2014 by McAuley et al. [9], by visiting a target book page and collecting the book recommendations that Amazon provided for that target. These relationships essentially list books that are, according to buyer behaviour, either substitutes for, or complementary to, the target book. The following relationships to a target book *A* are present in the dataset:

- buyers who bought *A also bought B*;
- buyers *bought together A* and *B*.

These relationships model a strong, undirected co-buying decision for both books involved, and amount to 33,058,487 co-buying relationships. The dataset did not include an essential piece of metadata for the books: the names of their authors. We thus completed the book metadata with author names using the public book OpenISBN records[1].

Related Work. Thelwall [11] ran recently a count of the gender of authors (classified based on first name, with ambiguous cases removed) per annotated writing genres, as reflected in a sample of 0.5 million books crawled from goodreads.com. It confirms substantial gender differences in authorship in some genres, such as romance and comics. Male authors are in the majority in most genres, except for children, adult, fantasy, suspense and cook books. However, the readership is close to gender-balanced in most genres, as far as the reviews on Goodreads can inform, if with a strong likely bias due to this website's 76% female user base. Krebs [5] is the pioneer text on the notion of web-mined book communities. Shi et al. [10] used the book co-buying relationship between science books and politically inclined books to indirectly compute the biased reading preferences of American readers across the political spectrum (also inspired by Krebs' seminal article [4]).

[1] www.openisbn.com/

2 Method

The main focus in the analysis of the 1,748,925 ISBN-assigned books is on their authors; in case of a list of authors, the first author in the sequence is taken to represent (assign gender to) the book. Effort is made to correctly classify the gender of each author using gender-annotated first-name datasets, plus manual lookups for the most frequent full names, pseudonyms, and authors with initials. After this gender classification, we investigate de facto gender mixing across the graph of tightly knit (*also bought*, and *bought together*) undirected book relations. For this, we compare two distributions: the observed gender of books co-bought with male-authored and female-authored books, respectively.

As this test strongly shows that the gender mixing is not neutral, but rather that there exists preferential gender association (female books are more likely to be co-bought with female books), we then dig deeper into the book graph to study whether gender disassociation is local to a writing genre, or ubiquitous. We segment the undirected book graph into book communities based on the graph structure alone (after a first step of unbiased random sampling the graph into a subgraph), visualize the communities obtained, and learn examples of book genres which are currently gender-polarized.

2.1 Classification of authors by gender

An author name is classified across four categories: *male* or *female* (when this gender could be unambiguously determined), *anonymous*, or otherwise is left *unknown*. Fig. 1 summarizes the classification method.

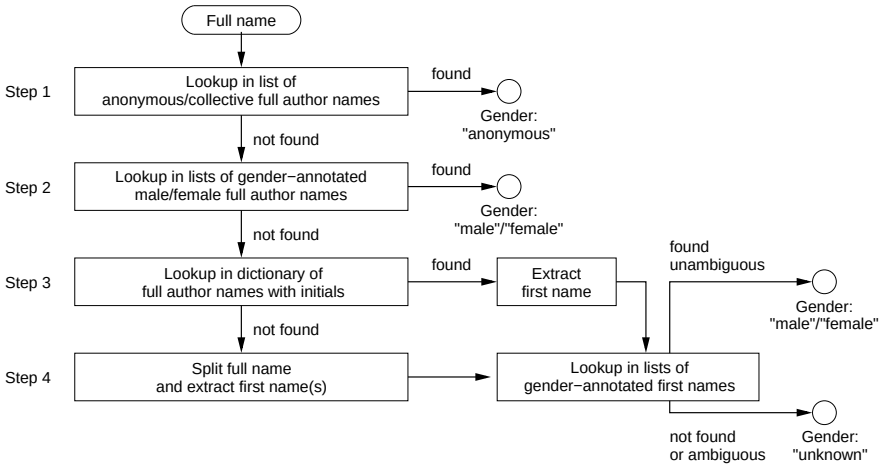

Fig. 1 Steps in determining the gender of an author's full name

Step 1. We first obtain a list of 8500 *anonymous* full author names present in the original dataset; this set consists of a majority of collective authors (such as the *Bodleian Library*, *Center For Constitutional Rights*, *Correspondents Of The New York Times*, *Creative Teaching Press*, *Editors At Scientific American*), and a minority of explicitly anonymous placeholders (*Unknown*, *Unknown Walker Author*). A book with no author specified at all is also classified as anonymous. This list is extracted from the author names present in the book dataset, by searching for keywords which are a giveaway for a collective or anonymous author (*Library*, *Press*, *Congress*, *Editors*, etc.) and then manually inspecting the entries selected to remove names not belonging to this category (e.g., *Andrea L. Press*, a female author in sociology). Then, all books whose authors remain in this list are classified as anonymous (Step 1 in Fig. 1). Due to the manual verification, this classification step is expected to be entirely accurate.

Step 4. Most of the remaining authors are classified by extracting from the author's full name their first name(s), and querying annotated lists of first-name use in the real population. This is done by Step 4 in Fig. 1. Three distinct data sources are used in this step:

(a) a corpus of annotated first names collected by the School of Computer Science at Carnegie Mellon University (CMU)[2], which categorizes approximately 5000 female and 3000 male first-name variations;
(b) a probabilistic library[3] 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;
(c) our own list of manually annotated first names from non-English-speaking countries, present among the authors whose books are sold on `amazon.com` (e.g., *Geneviève*, *Oana*).

Step 2. Step 4 by itself is not foolproof, as there exist cross-gender pen names (*George Eliot*, the pseudonym of the female writer Mary Anne Evans), as well as ambiguous first names: over 300 first names from the CMU corpus are used by either gender (e.g., *Andy*, *Page*, *Dana*). For this reason, Step 2 precedes it: it attempts to categorize the full names of the most popular authors in the dataset who have ambiguous first names, by relying on surnames to make a difference.

For this, we manually search Amazon and the wider Internet for author home pages, or any other concrete indication of the gender these authors subscribe to; these results make up new annotated lists of full author names (approximately 650 female and 900 male; e.g., *Stacy Phillips* is the male author of books in the music genre, while *Stacy Gregg* is the female author of children's books on horses). This list includes the authors who use initials, have never made public their first names, but whose gender is known (e.g., *C. J. Carmichael*, a bestselling female romance author). It also contains a small number of collective or anonymous authors whose gender is clearly predominantly female (e.g., *Asian Women United of California*) or male (e.g., *A Monk of the Eastern Church*). Note that, although constructed man-

[2] `www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/`

[3] `www.github.com/malev/gender-detector`

ually, these annotated lists of author names 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.

Step 3. Other authors do use initials (either always or occasionally), but their first names have also become publicly known (e.g., the poet *E. E. Cummings* is *Edward Estlin*). 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 350 names. Step 3 attempts to resolve an author name by looking it up in this dictionary; if found, then the first name is passed to Step 4 to attempt gender classification. The same theoretical limitation as for Step 2 holds: two published authors with identical initialed names may exist, in which case we have aggregated both into the name of the most popular author.

The annotated lists of full author names (Step 2) and the dictionary of full author names with initials (Step 3) could still be completed further; the author names which we manually classified belong to the most prolific and connected authors in this Amazon dataset. Nevertheless, at the end of this 4-step procedure, a large fraction of the author names in the dataset is expected to be accurately classified, with the remaining authors left unknown.

2.2 Statistics on author mixing by gender

Studying gender mixing amounts to finding whether or not readers are equally likely to buy female books in conjunction with male books, as in conjunction with female books. A *neutrally mixed* book market would mean that male books associate with (are direct neighbours of) female books at the same rate at which female books associate with female books; this does not imply that the books need be equally split between genders. A *disassociative* book market would see a gender preferentially neighbour the same gender.

To investigate whether genders mix or not, we first define the concept of *femininity of a book's neighbourhood*. Given any book b, this is *the fraction of female books* in b's direct neighbourhood, regardless of the size of the neighbourhood. Any value to this metric will fall in the normalized interval $[0, 1]$. Importantly, this neighbourhood excludes books with the *same first author* as b, to remove that association bias.

We then calculate and compare two samples:

(a) the femininity of the neighbourhood of male books;
(b) the femininity of the neighbourhood of female books.

These samples may be unequal in size.

A Kolmogorov-Smirnov test compares the two samples; its D statistic gives a numerical distance between the empirical distribution functions of the two samples, with the null hypothesis that the samples are drawn from *the same distribution*. If

the null hypothesis is rejected by the test, the book market is empirically found to lack neutral gender mixing. We use the Kolmogorov-Smirnov implementation in the Python *scipy* library.

2.3 Graph sampling and structural community detection

Independently of running the test for sample comparison, zooming into the local structure of the undirected book graph by segmenting it structurally into empirical *book communities* shows whether there exists local gender disassociativity at the community level. The assumption is that structural book communities will be found to roughly correspond to book genres or topics.

Structural community detection. The problem of community detection loosely aims at outputting, as communities, subgraphs so that the method maximizes the modularity metric within these graph divisions, i.e., their density of edges. In this study, it suffices to define book communities as non-overlapping, and no weights are assigned to the edges in the book graph. The problem is complex, as the simpler clique-finding problem is already NP-complete [2].

In a comparative study on the performance of 12 community-detection algorithms tested over a variety of graph structures [6], the *multilevel modularity optimization* community-detection algorithm from Blondel et al. [1] ranks among the top three algorithms, with the added advantage of being among the top two in speed of execution, with a low computational complexity expectedly linear in the number of edges in the graph. This allows it to deal with graphs larger than 10^4 nodes; we use the Python implementation in the *igraph* library. The technique is multistep: it first computes small communities by optimizing the local modularity in the neighborhood of each node. These small communities are then modelled by supernodes, with the original graph becoming a smaller, weighted graph. The process repeats until modularity cannot increase.

Graph sampling. As we intend to visualize the book communities obtained, before structural community detection, given that the book graph is large (on the order of 10^6 nodes and 10^7 edges), we take the first step of randomly sampling the graph into a scaled-down subgraph by one order of magnitude, using an unbiased sampling method which is expected to roughly preserve, if scaled down, the graph properties. Of the numerous existing random graph-sampling methods, we selected the computationally efficient *random-node* (RN) method [7], which requires no parameter configuration: given a target size n (number of nodes) for the graph sample (here, 10% of the original), it randomly selects a subset of n nodes from the original large graph, and preserves only those edges with endpoints in the node subset.

3 Results

1,196,676 of the total of 1,748,925 books are connected to other books. A fraction of the connected books have relationships where the other endpoint is a book whose metadata is not in the dataset; these relationships are removed. The resulting undirected graph has 17,489,263 edges, a maximum node degree of 5737, an average node degree of 14.61, and the assortativity coefficient by node degree $r = -0.0317$, which signals a non-assortative graph[4].

3.1 Gender classification. Statistics on gender mixing

Gender classification. After applying the gender classifier over the connected books, 12.78% (152,932 books) could not be gender-classified. This is largely due to the expensive manual gender categorization required for the fraction of author names which are gender-ambiguous, and to a lesser extent due to the book metadata being web-crawled—the free-form fields in the dataset (such as the author names) are occasionally misspelled or inconsistently formatted, and could not be resolved programmatically into a gender.

87.22% of the connected books (1,043,744 items) could be gender-classified. Among the first authors of the gender-classified books:

- 634,446 (60.79%) are male,
- 348,670 (33.41%) are female;
- the remaining 60,628 (5.81%) are anonymous.

The first authors of the unconnected books are 58.27% male, and 36.13% female; this gives a slightly higher likelihood for books authored by females to remain isolated from the larger book community.

Statistics on gender mixing. When calculating the metric describing the femininity of the neighbourhood, for each book in the gender-classified book set, the removal of all books with the same first author from a book's neighbourhood led to a fraction of these books having no neighbourhood left. These are not included in the Kolmogorov-Smirnov test.

Figure 2 then visualises the complete samples as histograms: the neighbourhood-femininity metric for male versus female books (Figure 2, left) and that for male versus anonymous books (Figure 2, right). The histograms show that neighbourhood-femininity values of 0, 1, and 0.5 are the most likely; this is natural, since this metric is a fraction where both the numerator and the denominator are relatively small natural numbers.

The visualisation shows qualitatively that the underlying distributions are dissimilar between male and female books (female books are more likely to form a

[4] This non-assortativity of Amazon book co-buying graphs is confirmed by the assortativity coefficients of other Amazon crawls public at `http://networkrepository.com/`.

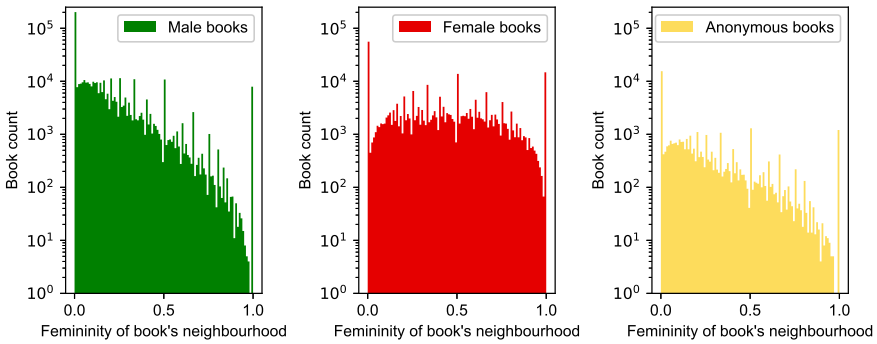

Fig. 2 The femininity of book neighbourhoods, i.e., for any given book b, the fraction of books co-bought with b which have female first authors. Across male books, the average femininity is 0.14; this value is 0.39 across female books, and 0.20 across anonymous books.

large fraction of the books co-bought with female books), and are instead relatively similar (if scaled down) between male and anonymous books.

Since the average neighbourhood-femininity metric is expected to be close to 0.33, due to there being fewer (33.41%) female books altogether, a neutral-mixing situation would see this neighbourhood-femininity value across the types of books. This is not the case in reality: while the average and median femininity metrics for *female books* are 0.39 and 0.37 (i.e., slightly higher than expected), the average and median femininity metrics are instead 0.14 and 0.06 for *male books* (i.e., substantially lower than expected), and are 0.20 and 0.11, respectively, for *anonymous books*.

The results of the Kolmogorov-Smirnov test comparing these samples of the neighbourhood-femininity metric are given by Table 1. With a low p value and a relatively high D statistic, the hypothesis that male and female books associate similarly can be rejected.

Table 1 Kolmogorov-Smirnov tests comparing pairs of samples of the femininity metric.

	D **statistic**	p **value**
male-female books	0.41	< 0.001
female-anonymous books	0.33	< 0.001
male-anonymous books	0.09	< 0.001

3.2 Graph sampling. Structural community detection and analysis

In order to visualise some of the local gender associativity in the connected book graph, a subgraph (of 119,667 vertices, or 10% of the 1,196,676 in the original connected graph) is sampled with the random-node method [7], and the resulting

sample graph is split into communities using the multilevel modularity optimization community-detection algorithm [1].

The largest 21 book communities are sorted by the percentage of books with female first authors, and are plotted in Figures 3 and 4 using the force-directed Fruchterman-Reingold layout implemented in the *igraph* library. Green denotes a book with a male, red a female, and yellow an anonymous first author. The books whose gender could not be classified are shown white. The communities have a wide range of gender compositions.

The communities shown vary from those dominated by male first authors (top in Figure 3) to those on the opposite side of the gender spectrum (bottom in Figure 4). They range in size from 611 books (female-dominated, bottom row left in Figure 4) to 3949 books (male-dominated, second row center in Figure 3), and have a percentage of female books ranging between 3.92% (top row, left in Figure 3) and 95.95% (bottom row, right, in Figure 4), calculated as a percentage among the books which could be gender-classified.

We thus observe strong gender preferential attachment, locally in some book communities. We attempt to classify the writing genres present in these 21 largest communities by randomly sampling 30 books from each community, and manually summarizing each sample. Table 2 gives these summary topics, in a grid corresponding to that of Figures 3 (the top five rows) and 4 (the bottom two rows).

fiction: fantasy, science fiction, thriller, comics, Marvel, DC Comics (3.92%)	**non-fiction**: Information Technology, logic, mathematics, applied science (7.58%)	**non-fiction**: music manuals, history of music, musician biographies (13.89%)
non-fiction: history, economy (15.29%)	**religious**: Christian, spiritual, Bible (18.88%)	**non-fiction**: economy, history, exact science, natural science, government (19.92%)
business, leadership, entrepreneurship, how-to (24.03%)	**non-fiction**: philosophy, metaphysics, history (27.07%)	**non-fiction**: encyclopedias, workbooks, nutrition, guides, how-to (29.89%)
alternative: spirituality, psychotherapy, occult, self-help (35.29%)	mainly **non-fiction**: biography, social, documentary, science, science fiction, classics, philosophy, education (35.36%)	**non-fiction**: books for health professionals (37.54%)
non-fiction: school manuals and practice books (38.07%)	**non-fiction**: therapy, mental health, self-development, learning (41.97%)	**fiction**: crime, mystery, adventure, thriller (43.96%)
non-fiction: cookbook, diet (49.54%)	**children's books**: fiction, nonfiction (54.13%)	**fiction**: historical, futuristic, fantasy, romance, science fiction (72.79%)
fiction: religious, historical, suspense, romance, fantasy, teen, female lead (73.17%)	**hobby**: decor, lifestyle, sew, quilt, knit, embroidery, design (81.89%)	**fiction**: romance, suspense, crime, adult, female lead (95.95%)

Table 2 Summary topics for the communities shown in Figures 3 (top five rows) and 4 (bottom two rows). The percentage of female books in each community is denoted in parentheses.

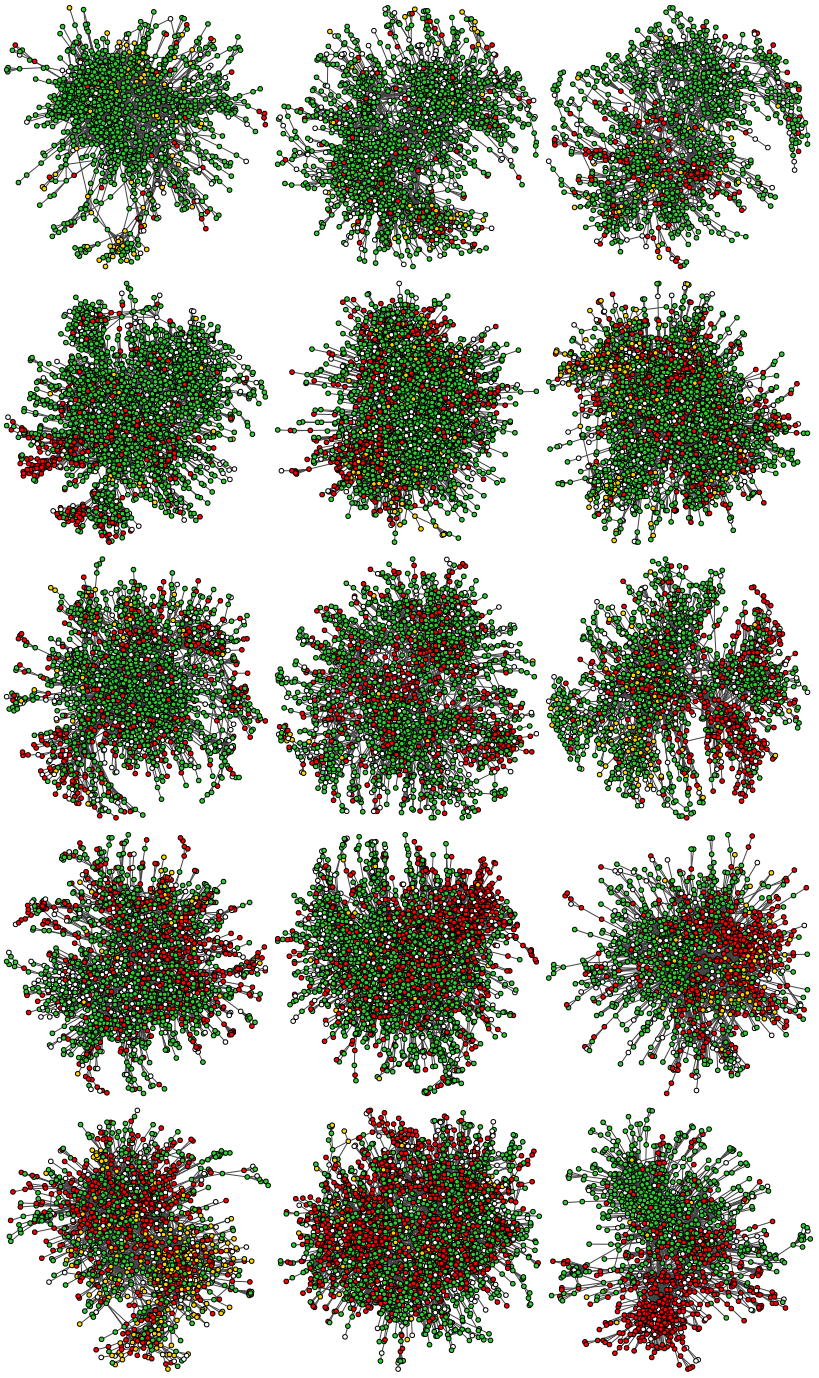

Fig. 3 Book communities (1). The colour of each book denotes the gender of the first author (green: male, red: female, yellow: anonymous, white: unknown).

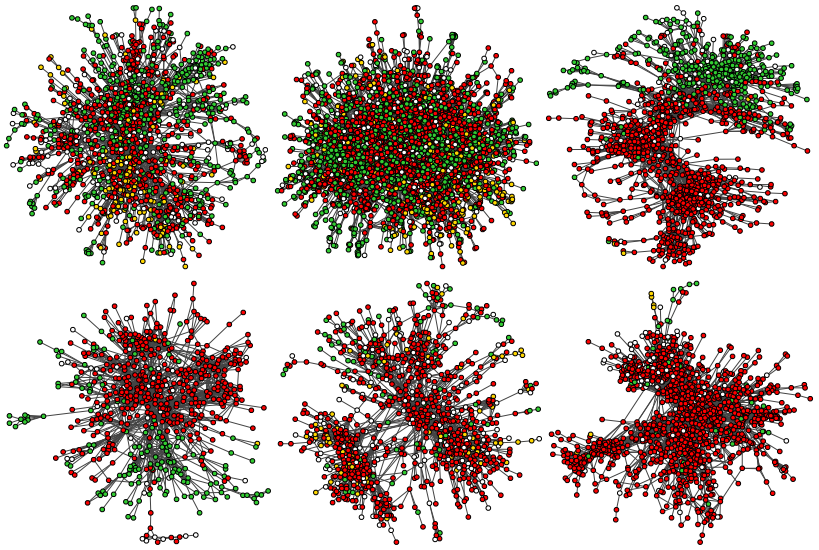

Fig. 4 Book communities (2). The colour of each book denotes the gender of the first author (green: male, red: female, yellow: anonymous, white: unknown).

This analysis of book communities gives two insights:

- Book communities are effectively *aggregated* book genres, grouped by buyer preferences (e.g., the non-fiction genres of biography, documentary, science, etc. are often consumed together).
- Book communities range from extremely polarised in terms of the gender of the authors to gender-balanced. The polarisation occurs roughly on the lines of the writing topics.

4 Conclusions

Summary of findings. Using a large dataset of `amazon.com` book co-purchases, we find empirically that author genders do not associate in a similar manner on the book market. Books by female first authors are significantly less likely than expected to be bought by readers who have chosen male authors before; this difference is *substantial*: the average femininity of the neighbourhood of a male-authored book is 0.14, which deviates from the expected 0.33 across the entire book market. In contrast, books by female first authors are only slightly more likely than expected to be bought by readers who have chosen female authors before: the average femininity of the neighbourhood of a female-authored book is 0.39, fairly close to the expected 0.33 value. This same-gender preferential association is local: certain writing topics are “coloured” preferentially by one gender.

While this gender assortativity can be attributed to more than one cause (a gender's preferential attachment to writing for one genre, or the buyers' preferential attachment to the output of writers of one gender), the end result are gendered book communities. This conclusion is similar to that of other recent findings: strong male-to-male academic prosociality in [8], and male-dominated collaboration structures in engineering in [3].

Limitations and discussion. The classification of author names by gender will have a small fraction of wrong calls, caused by the automated classification step using the first name of the author: it is possible that an author of one gender has a first name that is considered to belong unambiguously to the other gender. Also, the cleanliness of the text fields in the original dataset may occasionally lead to wrong gender calls; there exist author names which are truncated, misspelled, or completely missing. A cleaner dataset and a more complete, manually annotated dictionary of author names would raise the quality of the gender classification.

Future work. This study covered the American book market, so any insights gained related only to that geographic region. Future work will attempt to verify the findings on the European national book markets.

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment (10), P10,008
2. Fortunato, S.: Community detection in graphs. Physics Reports (3), 75 – 174. DOI 10.1016/j.physrep.2009.11.002
3. Ghiasi, G., Larivire, V., Sugimoto, C.R.: On the compliance of women engineers with a gendered scientific system. PLOS ONE **10**(12), 1–19 (2016). DOI 10.1371/journal.pone.0145931. URL `https://doi.org/10.1371/journal.pone.0145931`
4. Krebs, V.: Divided we stand??? (2003). `http://orgnet.com/leftright.html`
5. Krebs, V.: The social life of books, visualizing communities of interest via purchase patterns on the WWW (2004). `http://orgnet.com/booknet.html`
6. Lancichinetti, A., Fortunato, S.: Community detection algorithms: A comparative analysis. Phys. Rev. E **80**, 056,117 (2009). DOI 10.1103/PhysRevE.80.056117
7. Leskovec, J., Faloutsos, C.: Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06, pp. 631–636 (2006)
8. Massen, J.J.M., Bauer, L., Spurny, B., Bugnyar, T., Kret, M.E.: Sharing of science is most likely among male scientists. Scientific Reports **7**(12927) (2017). DOI 10.1038/s41598-017-13491-0
9. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: 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, pp. 43–52 (2015)
10. Shi, F., Shi, Y., Dokshin, F.A., Evans, J.A., Macy, M.W.: Millions of online book co-purchases reveal partisan differences in the consumption of science. Nature Human Behaviour p. 0079
11. Thelwall, M.: 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)