Exploratory Literature Review on Free Products: A Structural Topic Model Approach

Emilia Decoene

Dissertation written under the supervision of professor Miguel Godinho de Matos

Dissertation submitted in partial fulfillment of requirements for the MSc in Management with Specialization in Strategy, Entrepreneurship, and Impact, at the Universidade Católica Portuguesa, January 2, 2023.

Exploratory Literature Review on Free Products: A Structural Topic Model Approach

Emilia Decoene *

January 2, 2023

Abstract

Free products and services have increased tremendously in the last decade as a marketing technique and business strategy. We can observe this trend in the academic literature as well. In this exploratory literature review, we are interested in the different themes of this research area and how they have changed through time. We create a structural topic model (STM) to identify the underlying themes. An STM allows us to incorporate metadata into the model. This way, we can observe how the topics evolve depending on the publication year of the article. The study includes 279 academic papers from 1976 until 2022. We survey an increasing trend of themes situated in a digital world, especially for freemium products, and research concerning online consumer behavior. We identify two categories to classify free products; free products used as a marketing technique and free products as part of a business strategy.

Keywords: free product, free sample, free trial, freemium, structural topic model, literature review, exploratory analysis

*I would like to express my sincere gratitude to my supervisor, Miguel Godinho de Matos, for the invaluable guidance and support throughout the course of this project. His insights, encouragement, and motivation have been instrumental in helping me complete this dissertation. This work has been partially supported by the Portuguese Foundation for Science and Technology through research grant PTDC/EGE-OGE/27968/2017.

Revisão da Literatura Exploratória de Produtos Gratuitos: Uma Abordagem de "Structural Topic Model”

Emilia Decoene *

2 Janeiro, 2023

Resumo

Os produtos e serviços gratuitos aumentaram tremendamente na última década como técnica de marketing e estratégia comercial. Podemos observar esta tendência também na literatura académica. Nesta revisão exploratória da literatura, estamos interessados nos diferentes temas desta área de investigação e em como eles mudaram ao longo do tempo. Criamos um modelo temático estrutural (STM) para identificar os temas subjacentes. Um STM permite-nos incorporar metadados no modelo. Desta forma, podemos observar como os temas evoluem em função do ano de publicação do artigo. O estudo inclui 279 artigos académicos desde 1976 até 2022. Inquirimos uma tendência crescente de temas situados num mundo digital, especialmente para produtos "freemium", e investigação sobre o comportamento dos consumidores em linha. Identificamos duas categorias para classificar produtos gratuitos; produtos gratuitos utilizados como técnica de marketing e produtos gratuitos como parte de uma estratégia empresarial.

Palavras-chave: produto grátis, amostra grátis, teste grátis, freemium, modelo estrutural do tópico, revisão de literatura, análise exploratória

*Gostaria de expressar a minha sincera gratidão ao meu supervisor, Miguel Godinho de Matos, pela orientação e apoio inestimáveis ao longo de todo este projecto. Os seus conhecimentos, encorajamento e motivação têm sido fundamentais para me ajudar a completar esta dissertação. Este estudo tem sido parcialmente apoiado pela Fundação Portuguesa para a Ciência e Tecnologia através da bolsa de investigação PTDC/EGE-OGE/27968/2017.

Contents

List of Figures

List of Tables

1 Introduction

In the field of marketing strategies, businesses frequently offer clients free product samples to educate them about a product's features and suitability for their needs prior to making a purchase. As a result, companies create brand identity and awareness (Jain et al., 1995). The ultimate goal of the samples is to increase sales, and this outcome has been observed by Lammers (1991) and Gedenk and Neslin (1999). The samples were historically distributed in stores or at home, attached to another good, or as a standalone product.

Free samples have expanded since the emergence of digital goods and services. These are often experience goods for which quality is hard to ascertain prior to consumption. This causes businesses to provide users with a limited period of time to use their software for free or with a limited product version. Companies utilize such techniques because they want to convince customers about the quality of their product (Wang and Özkan-Seely, 2018).

The possibility of product sampling evolved into several different business strategies. For instance, Spotify's business objective is to persuade people to pay for its premium content while giving them access to the free version. In the literature, this strategy has been described as the "Freemium" strategy (Gu et al., 2018). Another example is observed at Apple Music, where customers have access to all content for one month. Apple Music's intention is to convince customers to continue using their product by signing up for a subscription plan. This strategy is sometimes referred to as a time-locked free trial (Dey et al., 2013).

Researchers have discovered new approaches and case studies to examine the application and efficacy of free sampling as a marketing technique in light of the significant shift in the industry from physical to digital goods. Since we are unaware of any recent research that have attempted to do a systematic review, we are interested in documenting this evolution.

In this paper, we provide an exploratory overview of this research field. We use machine learning tools to automatically organize and structure a large corpus through topic modeling. Utilizing this type of natural language processing, it is possible to automatically generate a list of the most relevant topics within the literature, along with the keywords associated with each topic.

We analyze 279 articles from highly acknowledged journals, published between 1976 and 2022.

We use these references to create a topic model that describes the literature on free products based on the abstracts of the surveyed papers. We describe the main topics on free products that we identify and we study how the different topics change through time.

We also examine the interrelation between topics, authors, and journals and determine whether we can classify topics into broader themes.

The structure of the review is as follows: in the literature review, we explain and examine the concept of topic models to find the best topic model for this type of exploratory research. Next, the methodology describes how the articles were collected and how it is converted to the input

data of the topic model. In this section, we also interpret the chosen topic model. The study of the articles and the various topics is the main emphasis of the "Results" section. Throughout this section, we identify the subjects, arrange them into a dendrogram, and examine how the topics change over time.

2 Literature Overview

Topic models are statistical methods that capture the main topics in a large set of documents by estimating which terms belong together over the different documents. It is an unsupervised machine learning technique, so we do not have to label the documents upfront. [1] We can use the models to see how topics are interconnected and how they change over time (Blei, 2012).

Topic models have proven to be an efficient way of describing what has been researched in a specific stream of academic literature.

The most common topic model is Latent Dirichlet Allocation, or LDA (Blei et al., 2003), and has been used before to find the underlying topics for exploratory literature reviews (Asmussen and Møller, 2020). Figure 1 shows a framework developed by Asmussen and Møller (2019) that defines the steps of using LDA for this specific use case.

The framework is divided into three steps: pre-processing, topic modeling, and post-processing.

The pre-processing step focuses on gathering the data, transforming it into the proper format, and choosing the parameters for the LDA model. The topic model is created in the second step and analyzed in the third step of the framework.

The post-processing of the topic model involves analyzing the topics, labeling them, and selecting the most important ones for the subsequent literature review. The framework provides an excellent base for an automatic literature review and can easily be extended.

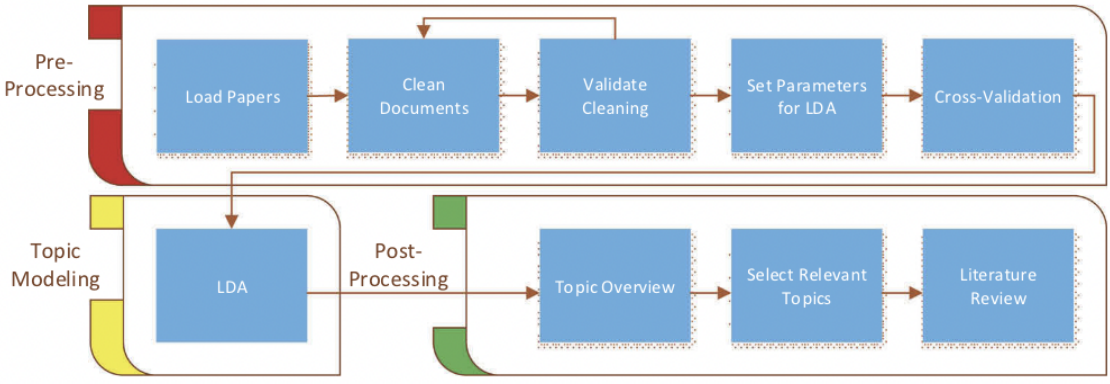


Figure 1: Overview of smart literature review framework (Asmussen and Møller, 2019)

[1] Unsupervised and supervised learning are the two main categories of machine learning. In supervised learning, the model retrieves the results of previously labeled examples to learn the results of new instances. We often apply this method to predictive analytics. Unsupervised learning involves the inference of data attributes without labels and is frequently used in descriptive analytics (Hastie et al., 2009).

LDA is a mixed-membership model, meaning that a document can be a mixture of several topics. This model contrasts with traditional models where one document belongs to one topic. A mixed membership model can provide a more accurate representation of the different academic interests in articles, which is helpful when searching subjects for future research.

Given that an LDA model is generative [2], it assumes that words in documents are produced by topics that disclose a hidden structure. Estimating the Dirichlet hyperparameters aims to uncover the documents' underlying patterns (Steyvers and Griffiths, 2007).

We describe an LDA topic model in the upper part of Figure 2. In the figure, D represents the set of all documents, and N represents the set of all terms in a document. The only observed variable is the n^{th} word $w_{d,n}$ in a particular document d. $z_{d,n}$ refers to the topic assignment of the n^{th} word in document d. Both $z_{d,n}$ and $w_{d,n}$ follow a Multinomial distribution.

LDA assumes that a topic exists of a distribution over terms. The model captures this through the variable β_k, referring to all term probabilities of the k^{th} topic. In total, the model holds K topics.

The second assumption in the model is that a document is a distribution over topics. The topic proportions in each document d are captured by θ_d. β_k and θ_d both follow a Dirichlet distribution, and α and η are the hyperparameters of the Dirichlet distributions (Blei, 2012).

We can estimate The LDA parameters in different ways. The most used algorithms are Gibbs sampling (Griffiths and Steyvers, 2004) and variational inference (Blei et al., 2003). [3]

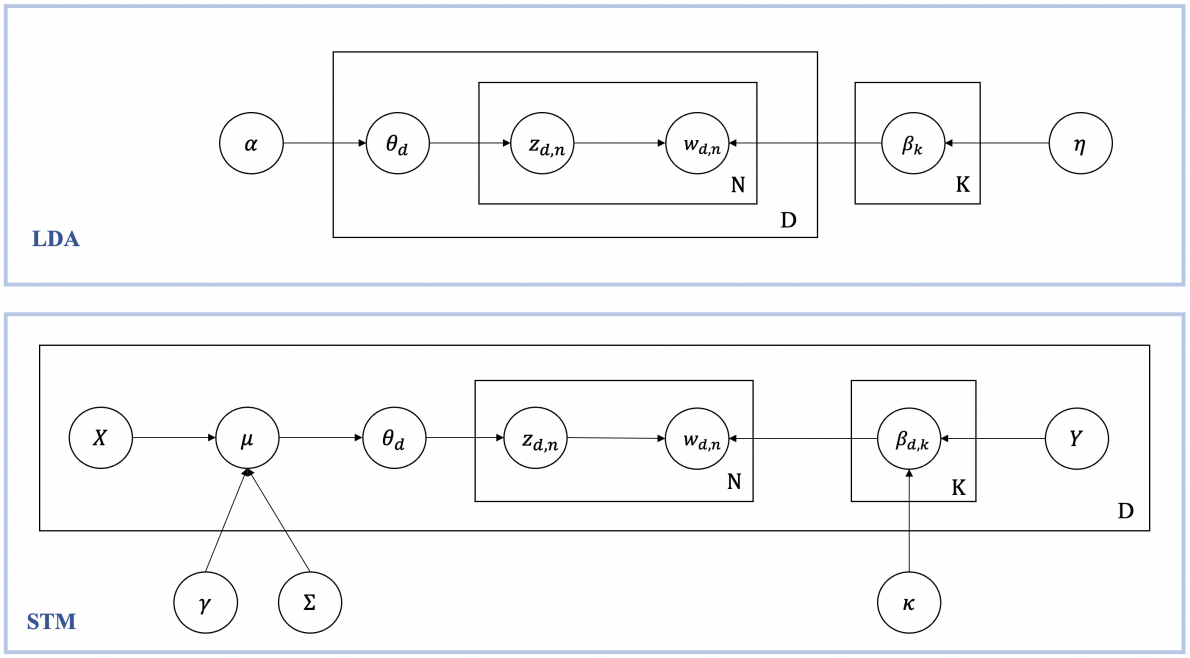


Figure 2: Notations of the LDA and STM model (Blei, 2012; Roberts et al., 2013).

[2] Models that can produce new data based on the data used to train them are known as "generative models (Foster, 2019)."

[3] Although we consider variational inference to be the faster solution, Gibbs sampling is often the preferred option when facing a smaller set of documents since the sampling method is easier to understand (Blei et al., 2017).

Albeit simple and easy to estimate, LDA topic models have certain drawbacks. For example, given the Dirichlet distributions in this model, LDA assumes that topics are independent of each other (Lafferty and Blei, 2005). However, when topics are surveyed within a particular research field, dependence and correlation between topics are reasonable assumptions, and it can be insightful to find out whether topics are related or if we can group them.

Another drawback of the model is that the topics are assumed to be independent of the documents and terms. We can see this in Figure 2 because the LDA model does not enclose β_k in the document collection D or the term collection N. This drawback also means that the model does not capture how topics can change because of covariates (Kuhn, 2018). The covariates refer to the characteristics of the documents, such as information about the author, journal, and publisher. We also call this the metadata of the document collection.

Covariates can have two different effects on topics. The distribution of topics in documents can change because of a certain covariate. For example, a certain topic can appear less in documents published in the '80s than nowadays. In this case, the publication year affects the topic's prevalence. This is called a topic prevalence covariate effect. Another example is that a topic has a higher probability of consisting of certain words because it was published in a particular journal rather than if it was published in another journal. In this example, the journal affects the topic's content, which is called a topical content covariate effect. Roberts et al. (2019) used this effect to describe the difference in a topic's word choice according to the political orientation of blog posts.

Unlike LDA topic models, a Structured Topic Model (STM) is another type of topic modeling that assumes a correlation between topics and incorporates the structure of a document collection by taking into account the metadata.

More specifically, in an STM model, the metadata is incorporated into the estimation process. The lower part of Figure 2 shows the model structure of an STM.

D and N represent the set of documents and set of terms per document, respectively. Each word $w_{d,n}$ is assigned to a certain topic $z_{d,n}$, both following the same distribution as in LDA.

The topic proportions in a document are still captured by θ_d but is now document-specific as it depends on $\mu = X_d\gamma$ and Σ. In the model, X_d is a vector of document covariates, while γ and Σ are a coefficients matrix and a covariance matrix, respectively. The distribution of θ is now Logistic-Normal and allows for correlation between topics (Roberts et al., 2014, 2019).

The distribution over terms is captured by $\beta_{d,k}$, which now also depends on the document. In Figure 2, we can see that $\beta_{d,k}$ is included in the document collection D, in contrast to the LDA model. The variable $\beta_{d,k}$ depends on the covariate vector Y that comprehends the content covariate. Instead of following a Dirichlet distribution, it follows:

$$\beta_{d,k} \propto exp(m + \kappa_k^{(t)} + \kappa_{y_d}^{(c)} + \kappa_{y_d,k}^{(i)})$$

A multinomial logit model is used to generate β, where m is the baseline log frequency for each possible word v in the lexicon. The deviations of m, measured by κ, are also included in the model of Roberts et al. (2013). These deviations refer to the deviation caused by the topic, the deviation caused by the covariate group, and the interaction of the two (Roberts et al., 2019).

STM also uses algorithms like Gibbs sampling and variational inference to estimate the topic models. To improve the estimation process, we can first employ the Spectral Initialization approach based on spectral decomposition to carry out the parameter initialization in an STM before we apply Gibbs sampling or variational inference. This approach makes the estimation process deterministic and consistent. (Roberts et al., 2019)

We are interested in studying how the research field of free products evolves, so we use an STM instead of LDA to estimate the topics. STMs have been used before to automate literature research, for example, in the field of maritime transport (Bai et al., 2021), and in social movements literature (Lindstedt, 2019).

3 Methodology

This section defines a topic model best suited for our research. In subsection 3.1, we analyze the data and build the topic model, and in subsection 3.2, we explain its properties. The analysis of the topics is discussed in the "Results" section.

3.1 Topic model pre-processing

Although topic models can accommodate a lot of data, the correct input format must be used before we can build the topic model. This implies that before the data is translated into the appropriate format, it must first be collected and cleaned. Since we are reviewing the literature on free products, the gathered data consists of publication information and abstracts of relevant academic papers. Before collecting the data, we conducted field research to find relevant keywords and areas to look for in the literature.

We can better understand the metadata by closely examining the journals and authors. We can gauge the scope of the research issue by building networks for author cooperations.

The cleaning process helps to reduce the data's complexity and noise, which makes modeling easier. This stage eliminates numbers and punctuation while testing the value of eliminating stop words and performing stemming. We select the best topic model as the final pre-processing step based on various evaluation indicators. The pre-processing stage is based on the method of Roberts et al. (2019).

3.1.1 Data collection

We collected the data by performing queries on the Scopus database. [4]

The queries searched for English journal articles or conference papers in the subject areas of business, social sciences, computer sciences, economics, and decision sciences. While Computer Sciences might seem like an odd area when looking for free products literature, this subject area also consists of digital business trends and we want to include digital products and business strategies.

In the Scopus database, we look for the following keywords: product sample/sampling, free service, free gifts, non-price promotion, zero price/pricing, free trial, free sample/sampling, free product, free content, product trial, free version, free to fee, paywall, free digital content, and freemium. The collection of these keywords is the result of preliminary field research.

The retrieved data consists of 2639 articles and contains information on the title, document type (journal article or conference paper), publication year, Scopus ID, DOI, journal, author, and the number of citations. Since Scopus can only find the first author's name, we created a different database via Mendeley by inserting the DOI numbers that return the names of all co-authors and are then matched to the article corpus. With this extra information, it is possible to analyze the researchers devoted to this topic in great detail.

From the collected database, we only keep the articles from highly acknowledged journals, according to the UT Dallas Ranking and the ABS list, or articles that belong to the 5% most cited. The percentage is arbitrarily chosen and excludes papers that were not published by the list of top journals and have less than 66 citations.

Because the database does not contain any information on the content of the articles, we fetch the abstracts from Scopus separately.

Next, we check all articles manually to see if they are relevant to the research scope. We eliminate certain articles in this step because they contain a keyword from the query but belong to an unrelated research field. For example, papers on public goods also appear in the database. In a way, these can be considered free products as well but we cleaned them because they are out of scope. Research on healthcare and insurance products also emerges in the database but should be left out for the same reason. Other papers referred to sampling in the statistical context, for example, in mathematics, economics, etc. This results in the removal of 111 articles because they belong to the area of Physics (6), Computer Science[5] (9), Mathematics (22), Healthcare Industry (21), Operations Research (13), Economics (17), Ecology (7), Public sector (7) and Other (9). [6] The final database consists of 279 articles and will be used to create the structural topic model.

[4] Scopus is Elsevier's largest database of academic articles

[5] Although we had previously included this category specifically, we discovered that several articles in this field did not fit the scope.

[6] All code, kept articles and manually removed articles including their appropriate research field can be found on https://github.com/emiliadecoene/exploratory-literature-review-free-products.

3.1.2 Metadata analysis

In addition to abstracts, the database includes other information on the articles that is helpful for researching this area of literature. This section analyzes the different journals and authors present in the data collection.

When we look at the journals incorporated in the database, we observe 95 distinct journals. In Figure 3, we list the 20 most popular journals in order of appearance, from highest to lowest. The International Journal of Research in Marketing has the most articles in the corpus, 12, followed by 11 publications, each from Management Science, Journal of Retailing, Journal of Marketing, and Journal of Management Information Systems. Seven of the 20 journals listed in the Figure are specifically about marketing. In the entire corpus, one journal appears four times, whereas nine journals appear three times, 14 journals appear twice, and 51 journals appear only once.

The database contains 600 unique authors. Figure 4 lists the authors having two or more publications in the data collection. [7] The figure does not distinguish the first authors from the co-authors. Professor Yong Tan appears to have the most publications (6), followed by Professor Juho Hamari (4). Eighteen authors wrote or co-authored three articles, while 43 wrote or co-authored two.

[7] We write the author names according to the regular expression *"Name[A-Z]∗Surname"*.

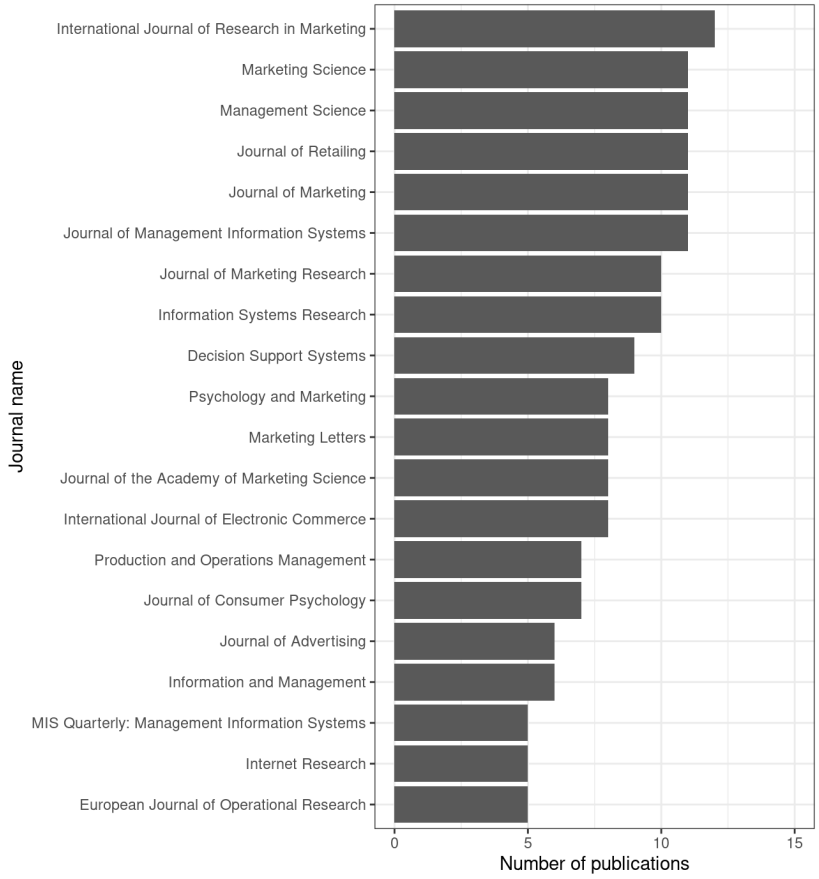


Figure 3: Most common journals

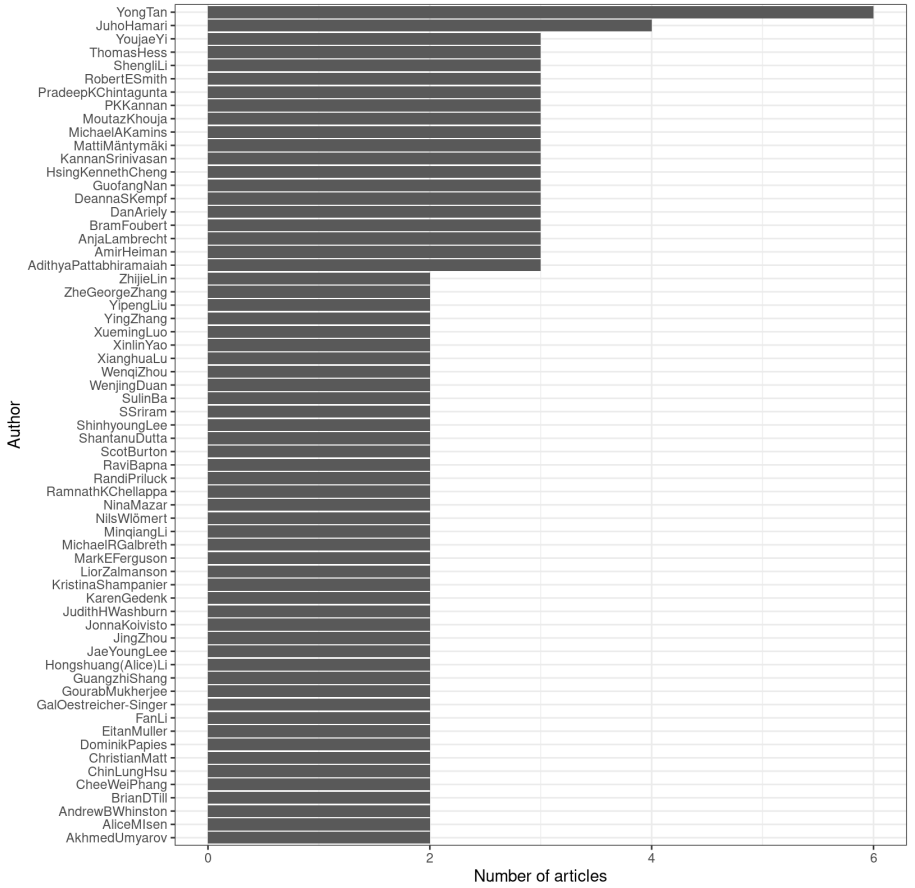


Figure 4: Authors with more than two publications

In order to have a better understanding of how the authors are related to each other, we include a graph model in Figure 5 that highlights the connections between authors. The network in the figure, which has 600 nodes and 745 relationships overall, gives a complete overview of all networks that result from various collaborations.

Nodes contain the author's name, and a relationship is drawn between nodes when authors co-write an article. Authors that published more than one article are highlighted in yellow. Note that authors that wrote an article without a co-author are not included in the graph.

If authors wrote an article that belongs to the 5% most cited publications in our database, the relationship between those authors is thickened. In this case, the publication has more than 65 citations. The relationships' colors denote the publication's time interval, as explained on the right side of the figure. If the same author group published several articles in the same time interval, only one edge is drawn between the two authors. This is because the graph overwrites relationships between the same authors. However, when one of the articles in the same time interval belongs to the highly cited literature and the other paper does not, we construct an additional relationship to capture both data elements in the graph.

Our analysis revealed that the majority of networks in this field are small, comprising only two or three authors. This suggests that the level of collaboration among authors is relatively low and that the research is widely dispersed. However, some larger networks can be seen in the upper left corner of Figure 5, which may be the result of authors working with a diverse group of co-authors on multiple papers or of a single publication that has many authors. Overall, the dispersed nature of the networks suggests that this literature field is being studied by many different research communities.

Figure 5: Networks of Author Collaborations

We zoom in on the graph networks that include authors with more than one publication. Appendix A shows 33 networks that result from the authors listed in Figure 4. Together, the networks contain 197 authors and 312 relationships.

On average, a network contains almost six nodes. The most extensive graph is network (e), having 13 nodes. Five authors in this network published more than one article in the corpus. The graphs are also helpful in analyzing whether authors were active in a certain period or multiple periods. For example, network (z) shows activity in 3 different periods, from the time interval "2000-2009" until "2020-2022".

3.1.3 Data cleaning and model setup

Since we are interested in the evolution of the literature on free goods, our model includes a topical prevalence covariate referring to the publication year of the article. We employ this approach in two ways and will compare the results to determine the best-performing model.

The first way of measuring time effects is by including the publication year variable. The alternative is to include a time interval as a covariate.[8]

Figure 6(a) describes the continuous time variable. It shows that the number of publications increased since 1994 and at an accelerated pace since 2010.

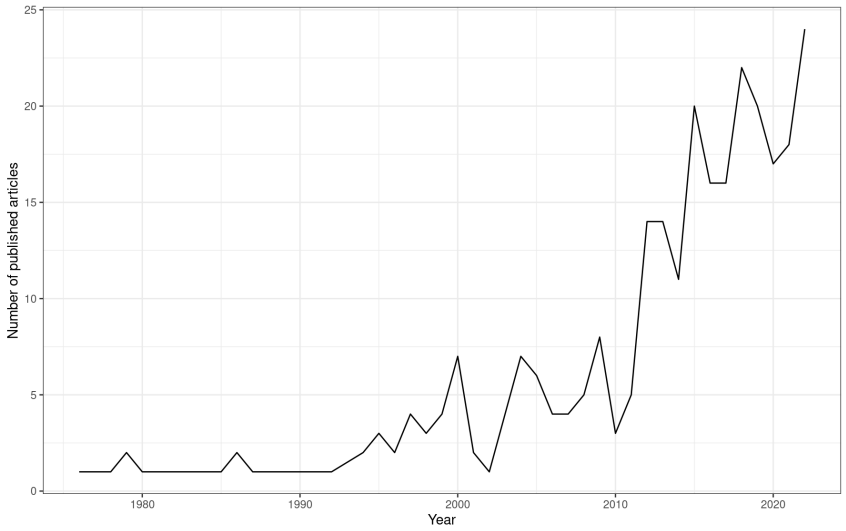
Figure 6(b) describes the discrete-time variable that we consider. Since we are dealing with unequal interval widths, we divide the width by the number of years per width. We observe that the yearly base is growing every time interval. In absolute terms, the interval "2010-2019" has the highest number of publications; 141. A total of 59 articles that meet our search are located in the years "2020–2022", 47 in "2000–2009", 21 in "1990–1999", and 11 earlier.

After gathering all the necessary variables for the model, we clean the data by removing digits and punctuation from the abstracts. We convert all characters to lowercase and remove words that occur in only one document.

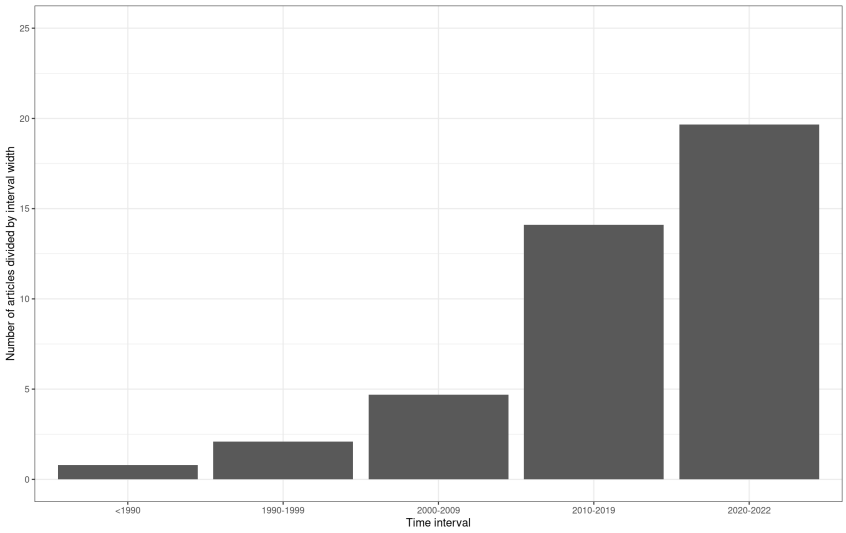
Because our document collection is small (279 documents) and consists of short texts, we want to ensure we do not omit essential words by stemming and removing stop words during the cleaning process. Also, several resources have shown that stemming and stop word removal do not significantly improve a topic model (Schofield and Mimno, 2016; Schofield et al., 2017).

Therefore, we create multiple document term matrices based on different data-cleaning processes. The document term matrix is the data input format for the topic model. It lists all abstract words in the matrix rows, and the numbers of the documents to which they belong are listed in the columns. The cells in the document-term matrix count how many times a particular word occurs in a document. Four different document-term matrices are created; one where the words are stemmed and stop words are removed, one where the words are not stemmed and stop words are removed, one where words are stemmed but stop words are not removed, and one where the words are not stemmed and stop words are not removed.

[8] We select time intervals as "<1990", "1990-1999", "2000-2009", "2010-2019", and "2020-2022". The intervals are uneven and they were chosen looking at how many articles were written over the years.



(a) Historical evolution of literature on free products



(b) Number of articles published per time interval on a yearly base

Figure 6: Number of publications according to two different time variables

Since we do not know how many underlying topics there are, we fit 28 models for each model setup, starting from 3 topics to 30. Roberts et al. (2019) recommend using spectral initialization to initialize the model parameters. After the parameter initialization, variational inference is applied. We choose the number of topics by observing that the optimal number lies somewhere in between after fitting for higher numbers of topics. Every topic model is evaluated using metrics such as exclusivity, semantic coherence, held-out log-likelihood, and residuals. The evaluation metrics are explained in Table 1.

Table 1: Explanations of evaluation metrics

Evaluation metric	Explanation
Exclusivity	Measures how exclusive a word is to a topic. A word is exclusive to one topic if the frequency of the word is low in other topics (Roberts et al., 2014).
Semantic Coherence	The degree to which terms in a topic are semantically meaningful to each other by measuring how frequent different words co-occur in that topic. (Mimno et al., 2011)
Held-out log-likelihood	Measures how well the model predicts words in unseen documents by comparing the log-likelihood of held-out documents with the log-likelihood of the full document collection (Wallach et al., 2009)
Residual analysis	A statistical technique used to assess the accuracy of a model. It involves examining the residuals, which are the differences between the observed values and the values predicted by the model. The lower the residuals, the more accurate the model. (Taddy, 2012)

Out of the 28 models, we narrow down the options by looking at the metrics and then choosing the best model by trading off the exclusivity and semantic coherence of the models (Roberts et al., 2019) for different K's. The metric plots can be found in Appendix B per setup. Interpreting the plots helps us choose the best topic model per setup. Summarily, there are eight different setups, depending on the four data cleaning methods and the two covariate options, and for every setup, we measure the optimal number of topics by fitting every setup 28 times. In total, we fit 224 (4x2x28) models to find the best topic model. Figure 7 outlines the different steps of this process.

After fitting every model setup 28 times, we compare the eight best models that are left. Table 2 summarises the metrics for the best models, each according to a different setup scenario. We divide the table in two, depending on which topical prevalence covariate we used: time intervals or the publication year. Every row in the table represents a different cleaning method. Finally, in the cells, we observe the number of topics K and the evaluation metric coefficients.

The model with the highest exclusivity of 9.18 has a publication year covariate, does stemming, and has removed stop words. The model with the highest semantic coherence of -47.87 has a time interval covariate and does no stemming nor stop word removal. The model with the highest held-out log-likelihood of -6.17 has a time interval covariate, removes stop words but does no stemming. Finally, the model with the lowest residual analysis outcome, 1.49, is the one with stemming and without stop word removal, having a time interval covariate. With

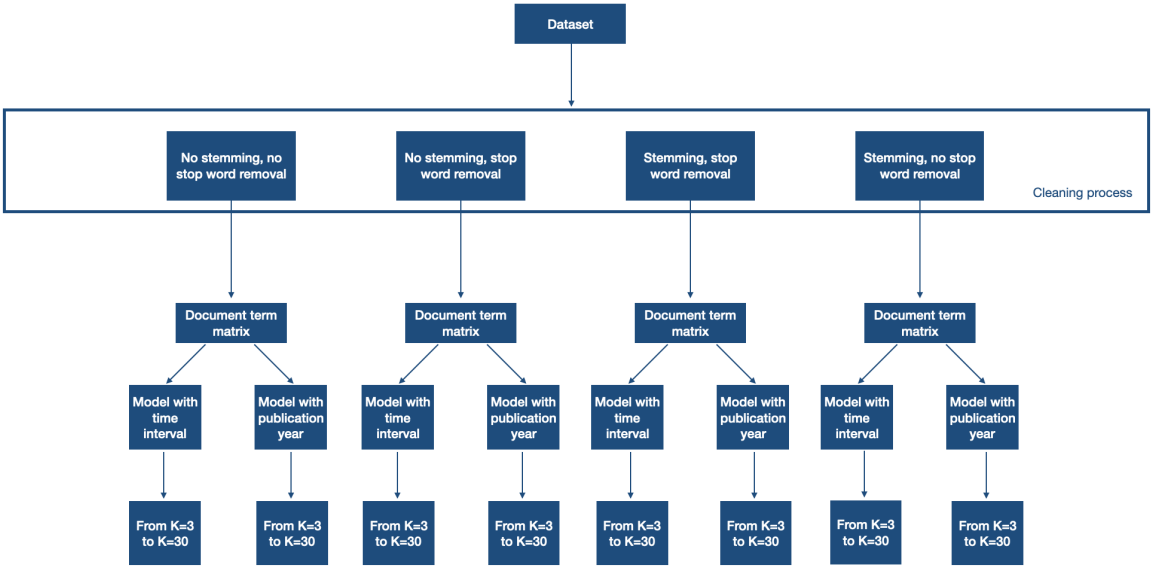


Figure 7: Overview of topic model pre-processing

a single evaluation metric in mind, we believe the aforementioned topic models to be the best. They consist of 13, 9, 12, and 11 topics, respectively. As we are most interested in obtaining a topic model with semantically coherent topics to understand the different aspects of the literature, we continue with the model with nine topics. The exclusivity of this model is 8.44, the held-out log-likelihood measures -6.73, and the residual analysis outcome totals 1.52.

Table 2: Overview of topic model metric coefficients

		Topical Prevalence Covariate									
		Time interval					Publication year				
		K	Exclusivity	Semantic coherence	Held-out log-likelihood	Residuals	K	Exclusivity	Semantic coherence	Held-out log-likelihood	Residuals
Data cleaning process	Stem & stop word removal	13	9.14	−69.85	−6.82	1.52	13	9.18	−71.28	−6.74	1.54
	Stem & no stop word removal	12	9.15	−80.39	−7.30	1.49	12	9.13	−81.88	−7.64	1.54
	No stem & no stop word removal	9	8.44	−47.87	−6.73	1.52	9	8.39	−49.12	−6.72	1.54
	No stem & stop word removal	11	8.75	−54.99	−6.17	1.52	11	8.66	−50.76	−6.23	1.52

3.2 Topic model

The topic model returns a matrix listing the topic-term probabilities (β-matrix) and a matrix that specifies the document-topic probabilities (γ-matrix). The β-matrix consists of 23,616 rows and three columns. Every row lists the probability that a term belongs to a topic. For example, the term 'online' has a probability of 0.0206 being generated by Topic 3, which is one

of the highest probabilities in the matrix. The following section will explain which terms are associated with certain topics to examine each topic's content. The γ-matrix counts 2511 rows and three columns and notes the probability that each topic is created in each document. Figure 8 displays the order of topic prevalences by grouping the γ-values of all topics appearing in the document collection and ranking the means. Topic 8 reaches the highest topic prevalence of 0.140, while Topic 9 comes in second with a prevalence of 0.127. This means that, on average, 14% of a document is generated by Topic 8, and 12.7% is generated by Topic 9. The prevalence of Topic 3 is the lowest (0.085).

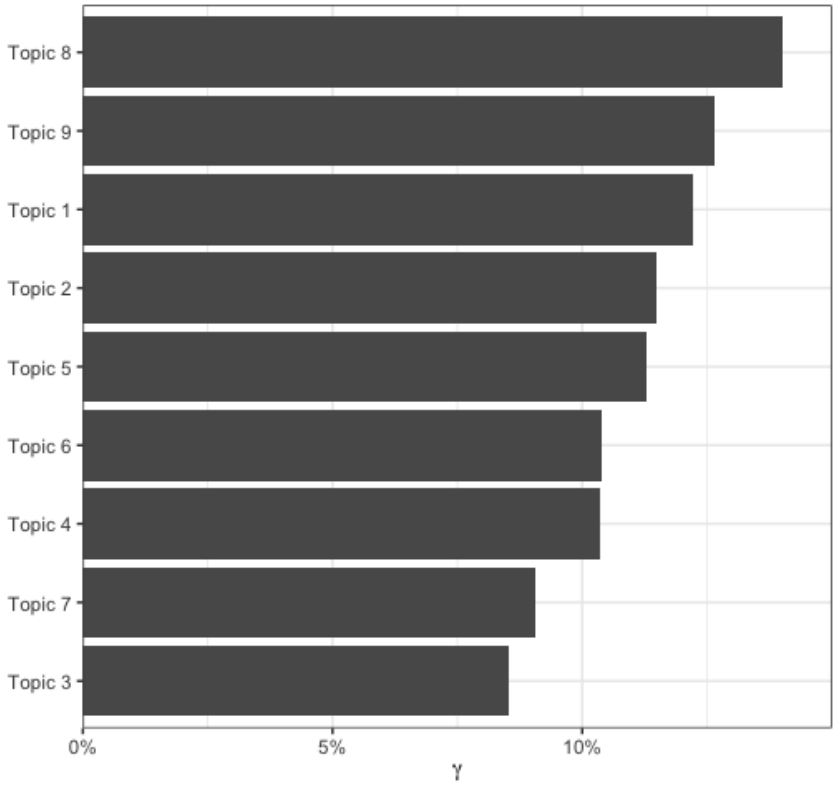


Figure 8: Topic prevalences

4 Results

4.1 Topic model analysis

In this section, we analyze the topics from the topic model. First, we select a label by examining each topic's most common terms and evaluating the papers with the highest probability of being generated by that topic. Since structural topic models build upon correlated topic models, we can investigate how topics are related to each other in a dendrogram. Finally, we analyze the effect of the time covariate on the prevalence of topics and examine which topics occur more in specific periods.

4.1.1 Topic inspection and labeling

To summarize the topics, we observe the ten terms with the highest FREX scores rather than the highest probability words (see Table 3). This is because there was no stemming nor stop word removal so the highest probability words would have shallow meaning. The FREX scores focus on the frequency and exclusivity of words in topics simultaneously (Roberts et al., 2013). In this subsection, the labeling process is clarified. The words in Table 3 are listed from highest to lowest FREX score.

Furthermore, we examine the documents most likely produced by each topic in question based on the γ-matrix. We characterize how the topic was used in those documents. By selecting the γ-values greater than 0.5, we can analyze the documents that belong mainly to one topic. Figure 9 displays the number of articles with a probability of higher than 0.5 of being appointed to a specific topic. Topic 8 has the most assigned documents (39), followed by Topic 9 (30). Topic 3 has the least assigned documents (21). Table 4 lists the documents with the highest γ-values per topic, its citation, and the γ-value. Together, the terms and document titles form the base to label a topic.

It should be taken into consideration that only the titles are mentioned in Table 4 while the topics are generated from the abstracts of the documents. Therefore the title may not always accurately reflect the topic of the paper. However, it provides a general idea of the focus of the paper. The explanations in this analysis are based on the first 10 FREX phrases and the first 5 document titles, which may not provide a comprehensive understanding of the semantic meaning of the topic. It is necessary to consider this limitation when interpreting the identified topics. Therefore, when categorizing the subject, we try to be as broad as possible but within the context of free products. The corresponding labels can be found in the last column of Table 3 and the first column of Table 4.[9]

[9] The complete document information, including the abstracts can be found in the database on https://github.com/emiliadecoene/exploratory-literature-review-free-products.

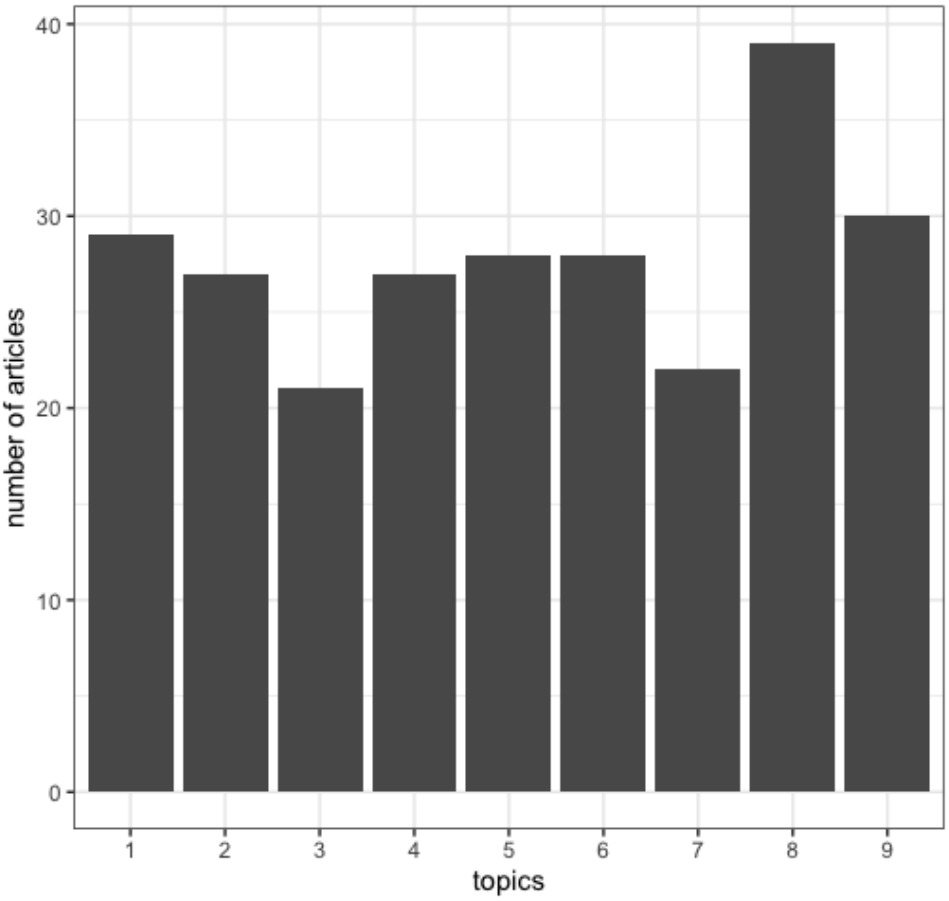


Figure 9: Number of articles appointed to a topic ($\gamma > 0.5$)

Table 3: Topic FREX words and labels

	Highest FREX words										**Topic label**
Topic 1	version	app	apps	cards	strategy	versions	freemium	news	paid	mobile	**Freemium Strategy**
Topic 2	reward	firm	firms	buyers	segment	concept	programs	software	seller	optimal	**Reward Programs**
Topic 3	wom	internal	coupons	loyalty	websites	online	targeted	gifts	sites	in-app	**Consumer Behavior**
Topic 4	social	players	money	users	premium	in-game	personal	features	pay	spending	**Online Communities**
Topic 5	free-trial	samples	usage	blog	sampling	marketing	bloggers	music	sample	experience	**Free Samples and Trials**
Topic 6	brand	brands	dependence	trial	attributes	state	equity	partner	process	habit	**Branding**
Topic 7	zero	zero-price	thinking	streaming	technologies	demonstration	adoption	price	costs	good	**Zero Price**
Topic 8	cps	fragmentation	equilibrium	internet	termination	providers	actions	revenue	service	fees	**Internet Companies**
Topic 9	promotion	retailers	promotions	return	returns	gift	involvement	framing	promoted	previews	**Retail Promotions**

Table 4: Documents with highest proportions per topic

Topic	Number	Title	Citation	γ-value
Topic 1 Freemium Strategy	324	Freemium pricing in digital games with virtual currency	(Meng et al., 2021)	0.9865
	122	On the monetization of mobile apps	(Appel et al., 2020)	0.9863
	355	The implications of offering free versions for the performance of paid mobile apps	(Arora et al., 2017)	0.9847
	361	Optimal freemium strategy for information goods in the presence of piracy	(Nan et al., 2018)	0.9832
	296	Managing the versioning decision over an app's lifetime	(Lee et al., 2021)	0.9820
Topic 2 Reward Programs	71	An approach for determining optimal product sampling for the diffusion of a new product	(Jain et al., 1995)	0.9914
	261	Concept testing with and without product trial	(Dickinson and Wilby, 1997)	0.9890
	329	Reward programs and tacit collusion	(Kim et al., 2001)	0.9888
	275	Utilizing public betas and free trials to launch a software product	(Mehra and Saha, 2018)	0.9869
	80	The benefits of information asymmetry: When to sell to informed customers?	(Bhargava and Chen, 2012)	0.9869
Topic 3 Consumer Behavior	241	The super engagers of freemium gamified services: using multimethod approach to examine why highly interactive customers become paying customers	(Gupta et al., 2022)	0.9903
	82	An empirical study of how third-party websites influence the feedback mechanism between online Word-of-Mouth and retail sales	(Zhou and Duan, 2015)	0.9887

Table 4: Documents with highest proportions per topic - continued

Topic	Number	Title	Citation	γ-value
	152	Comparing the effectiveness of rewards and individually targeted coupons in loyalty programs	(Gabel and Guhl, 2022)	0.9855
	84	The interplay between free sampling and word-of-mouth in the online software market	(Chen et al., 2017)	0.9854
	242	Effects of free gifts with purchase on online purchase satisfaction: the moderating role of uncertainty	(Zhu et al., 2015)	0.9820
Topic 4 Online Communities	336	Do your online friends make you pay?A randomized field experiment on peer influence in online social networks	(Bapna and Umyarov, 2015)	0.9885
	75	Why do players buy in-game content? An empirical study on concrete purchase motivations	(Hamari et al., 2017)	0.9882
	178	Towards a value theory for personal data	(Spiekermann and Korunovska, 2017)	0.9869
	200	Uses and gratifications of Pokémon Go: Why do people play mobile location-based augmented reality games?	(Hamari et al., 2019)	0.9865
	198	How synchronous participation affects the willingness to subscribe to social live streaming:the role of co-interactive behavior on Twitch	(Bründl et al., 2022)	0.9863
Topic 5 Free Samples and Trials	234	The effect of community identification on attitude and intention toward a blogging community	(Shen and Chiou, 2009)	0.9887
	235	The effects of blogger recommendations on customers' online shopping behavior	(Hsu et al., 2013)	0.9867

Table 4: Documents with highest proportions per topic - continued

Topic	Number	Title	Citation	γ-value
	237	Optimizing product trials by eliciting flow states: the enabling roles of curiosity,openness and information valence	(Lavoie and Main, 2022)	0.9867
	39	A model of the role of free drug samples in physicians' prescriptions decisions	(Chen et al., 2017)	0.9842
	278	The signaling effect of sampling size in physical goods sampling via online channels	(Liu et al., 2022)	0.9842
Topic 6 Branding	327	A framework for investigating "the hand of the past", and heterogeneity of dynamic brand choice	(Roy et al., 1996)	0.9956
	1	Attitude formation from product trial: distinct roles of cognition and affect for hedonic and functional products	(Kempf, 1999)	0.9928
	328	A dynamic analysis of market structure based on panel data	(Erdem, 1996)	0.9913
	371	Consumer processing of product trial and the influence of prior advertising: a structural modeling approach	(Kempf and Smith, 1998)	0.9871
	5	Brand alliance and customer-based brand equity	(Washburn et al., 2004)	0.9871
Topic 7 Zero Price	333	Self-signaling and prosocial behavior: A cause marketing experiment	(Dubé et al., 2017)	0.9885
	42	Battle royal: zero-price effects vs relative vs referent thinking	(Nicolau, 2012)	0.9878
	81	Reliability (or "lack thereof") of on-line preference revelation: A controlled experimental analysis	(Chen et al., 2013)	0.9860
	111	Consignment contracts with cooperative programs and price discounts mechanisms in a dynamic supply chain	(Buratto et al., 2019)	0.9846

Table 4: Documents with highest proportions per topic - continued

Topic	Number	Title	Citation	γ-value
	326	Zero as a special price: the true value of free products	(Shampanier et al., 2007)	0.9843
Topic 8 Internet Companies	281	Improved retention analysis in freemium role-playing games by jointly modelling players' motivation, progression and churn	(Karmakar et al., 2022)	0.9879
	362	Advertising vs brokerage model for online trading platforms	(Chen et al., 2016)	0.9879
	92	Scheduling advertisements on a web page to maximize revenue	(Kumar et al., 2006)	0.9875
	318	Net neutrality, exclusivity contracts and internet fragmentation	(Kourandi et al., 2015)	0.9870
	349	Incentivized actions in freemium games	(Sheng et al., 2022)	0.9870
Topic 9 Retail Promotions	239	When free gifts hurt the promoted product: the influence of product-gift fit on activating persuasion knowledge and devaluating the promoted product	(Park and Yi, 2019)	0.9903
	115	The effects of promotional frames of sales packages on perceived price increases and repurchase intentions	(**?**)	0.9867
	265	Where should I focus my return reduction efforts? Empirical guidance for retailers	(Shang et al., 2019)	0.9865
	299	Hiding gifts behind the veil of vouchers: on the effect of gift vouchers vs directed gifts in conditional promotions	(Ding and Zhang, 2020)	0.9864
	8	"Seize the deal, or return it losing your free gift": the effect of a gift-with-purchase promotion on product return intention	(Lee and Yi, 2017)	0.9856

Topic 1. The most frequent and exclusive words in Topic 1 are "version", "app", "apps", "cards", "strategy", "versions", "freemium", "news", "paid", and "mobile". The words "version," "app," "apps," "strategy," "versions," "freemium," "paid," and "mobile" are all keywords of the freemium business strategy, which is frequently used to sell digital products and services such as mobile applications, games, and online streaming services. The freemium strategy offers a free digital product to customers, and the goal is to convince the customer to pay for the premium or paid version, often in the form of a monthly subscription. The γ-values of documents 324, 122, 355, 361, and 296 show the highest proportions of Topic 1. The recurring subjects in the titles are pricing and strategic decisions in a freemium business model. For example, document 324 deals with pricing in the context of digital games, while documents 122 and 355 apply to mobile apps. Document 361 and 296 refer to other strategic aspects of "Freemium", such as versioning. Looking at the research questions of these documents, they investigate the decision-making of monetizing apps and the implications of offering a free version. We label this topic as **Freemium Strategy**.

Topic 2. Topic 2 consists of the words "reward," "firm," "firms," "buyers," "segment," "concept," "programs," "software," "seller," and "optimal." These terms can be grouped into the concept of reward programs. Reward programs are programs where customers are rewarded for being loyal or trying new products. An example of a reward could be a gift or a sample of the product they are trying. The programs are a popular tool for physical products, but the term "software" indicates that they also exist in the digital world. The documents with the highest proportions of Topic 2 are 71, 261, 329, 275, and 80. Documents 71, 261, and 275 are examples of reward programs used for concept testing new products in the physical and digital world. The authors investigate the optimal circumstances for introducing these concept products. Document 329 uses the concept of reward programs in the context of loyalty programs and researches its impact on pricing. Document 80 investigates a broader scope, i.e., the benefits of not revealing information about a product before a purchase opportunity and its implications on a company's profit. The label of topic 2 is **Reward Programs**.

Topic 3. In Topic 3, we observe words such as "wom," "internal," "coupons," "loyalty," "websites," "online," "targeted," "gifts," "sites," and "in-app." The term "wom" refers to "word of mouth," where product users inform others about a particular product or service, either in person or digitally through online reviews and social media platform. The terms "coupons," "loyalty," and "gifts" apply to methods to increase loyalty among customers. Because words such as "websites," "online," "sites," and "in-app" are highly prevalent as well, we can assume that this topic is also categorized in the digital world. Looking at documents 241, 82, 152, 84, and 242, we observe different subjects, such as customer engagement, word-of-mouth, loyalty programs, and customer satisfaction. Documents 241, 82, 152, and 242 specifically mention online environments. In documents 241 and 242, the authors investigate the online behavior of gamers,

while in documents 82 and 84, the researchers examine the effects of online word of mouth in combination with free samples. This topic covers different facets of influencing consumer behavior. Since we can situate four terms and four documents in the digital world, we label this topic as **Online Consumer Behavior**.

Topic 4. The highest FREX words of Topic 4 are "social," "players," "money," "users," "premium," "in-game," "personal," "features," "pay," and "spending." The terms "social," "players," "users," "premium," "in-game," and "features" together apply to the context of online gaming. The documents with the highest probabilities of Topic 4 are 336, 75, 178, 200, and 198. Documents 336 and 198 cover social networking and the effects of peers and social interactions, while documents 75 and 200 focus on online gaming. These documents investigate whether socializing and other variables affect usage and purchase intentions. Document 178 discusses personal data and how online users value it, which is a more broadly based subject. In online business models, personal data can often be seen as the underlying currency in social networks, so this is an essential aspect of the topic. We categorize this topic as **Online communities**.

Topic 5. The terms in Topic 5 with the highest FREX scores are "free-trial," "samples," "usage," "blog," "sampling," "marketing," "bloggers," "music," "sample," and "experience." Words such as "free-trial," "samples," "sampling," "marketing," "experience," and "sample" refer to the usage of trials and samples as marketing tools. The goal of this tool is to convince potential customers. This way, they learn and experience the product's features before purchasing. The other terms, "blog," "bloggers," and "music," are more specific and might indicate different contexts in which this marketing technique is applied. The topic is most observed in documents 234, 235, 237, 39, and 278. The effectiveness of free samples and trials of physical products are discussed in documents 235, 39, and 278. Document 235 explores how a trial experience can be optimized, while Document 278 looks into how a sample size works as a quality signal. In document 39, the role of free samples is discussed in a different context, pharmaceutics. The titles of documents 234 and 235 are particularly about blogging, which is, in most cases, a free service. In these documents, the authors investigate the motivations behind using blogs as a free service and how it affects purchase decisions. In order to stay general, Topic 5 is labeled as **Free Samples and Trials**.

Topic 6. In Topic 6, the words are "brand," "brands," "dependence," "trial," "attributes," "state," "equity," "partner," "process," and "habit." These terms introduce the concept of "branding," which is the label of this topic. Brand equity can be built through free trials and partnerships with other companies. Customers acquire opinions about the various brand features through a cognitive process. Trials can be used to shape consumer perception of a brand. These concepts are captured in documents 327, 1, 328, 371, and 5. General branding concepts, such as brand choice, brand positioning, and brand equity, are mentioned respectively in documents

327, 328, and 5. Documents 327 and 328 both develop a model that incorporates dynamic brand choices and document 5 studies the effects of brand equity. The titles of articles 1 and 371 emphasize the thought process involved in product trials, as explained before. In document 1, the researchers examine the effect of brand cognition and other variables on free trial perception, and document 264 focuses on the trial process itself. We can group these ideas under the category **Branding**.

Topic 7. The highest FREX terms of Topic 7 are "zero," "zero-pricing," "thinking," "streaming," "technologies," "demonstration," "adoption," "price," "costs," and "good." Zero pricing is the pricing mechanism behind free products and has an increased effect on the expected utility function of customers, as opposed to non-zero prices (Shampanier et al., 2007). This technique can lead to faster product adoption. Words such as "streaming," "technologies," and "demonstration" are more difficult to put in this context. It is also less obvious what the topic is about when reading the titles of the high-probability documents (333, 42, 81, 111, and 326). The effects of zero pricing is the main research subject of documents 42 and 326. Document 111 covers price discounts and how it helps coordinating a supply chain. Although less obvious, the concepts in documents 333 and 81 also relate to price reductions and zero pricing upon closer examination, although it's not the main research topic. The authors investigate the effects of pro-social behavior on consumption choice and reliability of online preferences, respectively. We choose to identify Topic 7 as **Zero Pricing**, as it is most prominent in the FREX words and is more suited in the context of the literature field.

Topic 8. The keywords in Topic 8 are "cps," "equilibrium," "internet," "termination," "providers," "actions," "revenue," "service," and "fees." In this context, " CPs " means "content providers." The keywords "providers," "internet," and "service" refer to the market of online service providers. The other terms are more difficult to group. The documents with the highest probability of being composed by this topic are 281, 362, 92, 318, and 349. Three subjects can be identified. Freemium games are covered in documents 281 and 349. The authors investigate how to optimize retention and revenue-generating content in these games. In documents 362 and 92, researchers discuss advertising models by analyzing when advertising models are an optimal business model and how to optimize advertisements within this model. In document 318, the authors cover the internet market by addressing issues like fragmentation and net neutrality and how they relate to each other in different models. All these subjects relate to the decision-making process of companies on the internet, and thus we label Topic 8 as **Internet Companies**.

Topic 9. Finally, the highest observed FREX words in Topic 9 are "promotion," "retailers," "promotions," "return," "returns," "gift," "involvement," "framing," "promoted," and "previews." The topic terms relate to free products as a promotion tool for retailers where the goal is to have as many returning customers and involvement after offering them a free gift. The term "fram-

ing" refers to how a product or service is presented to potential customers. Companies can, for example, present price reductions differently to make them more attractive to the consumer. We observe different themes when looking at the titles of documents 239, 115, 265, 299, and 8. In papers 239, 299, and 8, the writers explore free gift promotions and examine how the gifts affect consumers' perceptions and intentions regarding the acquired product. The authors of document 115 investigate the effect of promotional framing and how it affects consumer price evaluations. Document 265 covers returns in retail and the effects of a trial length. These subjects can be bundled under the label **Retail Promotions**.

4.1.2 Topic dendrogram

Since STMs allow for correlated topics, it is possible to create a dendrogram that visualizes how the topics are related to each other (see Figure 10). A dendrogram is a graphical representation of a hierarchical structure, and it is constructed by clustering the topics based on the semantic coherence between topics. The clustering algorithm typically draws the branches of the dendrogram such that the vertical distance between the clusters is proportional to the distance in semantic coherence between clusters. Topics that are clustered at a lower level on the Figure have a higher coherence score.

Starting on the left side of the Figure at the bottom, Topic 6 ("Branding") and Topic 9 ("Retail Promotions") are grouped together. On the same level, Topic 2 ("Reward Programs") and Topic 7 ("Zero Price") are clustered. The algorithm merges the four topics with Topic 5 ("Free Samples and Trials") at a higher level or lower semantic coherence score. There is no differentiation between physical and digital settings on the left side of the dendrogram. As we've seen in the terminology and documents related to these topics, both contexts are included in these topics. However, all topics on the left side are examples of marketing techniques.

On the right side of the dendrogram, the algorithm joints Topic 3 ("Online Consumer Behavior") and Topic 4 ("Online Games") at the lowest level. Next, they are clustered with Topic 1 ("Freemium Strategy") and Topic 8 ("Internet Companies") at the lowest level of semantic coherence in the dendrogram. Finally, we can group all topics together as one. The right side of Figure 10 merely focuses on the digital environment from a strategic viewpoint as well as from a marketing viewpoint.

4.1.3 Topic prevalence covariate analysis

A structural topic model allows us to perform a trend analysis of the topics. In this subsection, we examine the relationships between the topics and the publication year of the documents. We are interested in which topics have a higher prevalence than others and how it changed over the years.

Table 5 outputs the regression results of the structural topic model. It shows how prevalent topics are in every time interval, starting from the baseline "<1990". Only Topic 4, Topic 5,

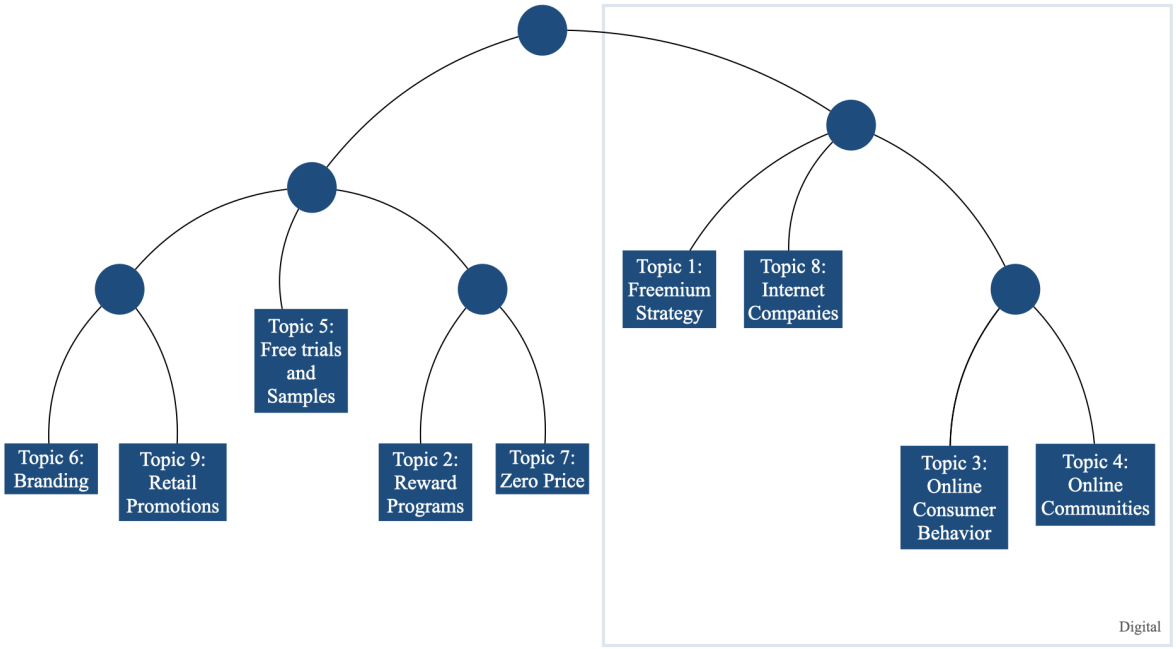


Figure 10: Topic model dendrogram

Topic 6, and Topic 9 indicate significant results. Regarding Topic 4, the mean topic prevalence during the time interval "<1990" is 0.176, given a 95% confidence interval. This means that the expected proportion of Topic 4 ("Online Communities") in a document written in the time period "<1990" is estimated to be 0.176. On a 10%-significance level, papers are expected to contain a propotion of 0.202 of information on Topic 5 ("Free Trials and Samples"), 0.169 on Topic 6 ("Branding"), and 0.150 on Topic 9 ("Retail Promotions") throughout the same time period. In the period of 1990-1999, the expected proportion of Topic 6 increased by 0.235 to the previous period, given a 95% confidence interval. It seems odd that Topic 4 ("Online Communities") has such a high prevalence in the time interval "<1990". This is due to the small number of articles (11) that were published during this interval compared to the other time periods. As a result, the regression findings will be strongly influenced if a document released during the time of "1990" contains a lot of terms that are highly prevalent in a particular topic. For this reason, the following paragraph will only cover the time periods after "<1990".

Structural topic models can also visualize the contrasts between different time periods. Figure 11 plots the differences in topic prevalence between the time intervals "1990-1999" and "2020-2022" with a 90%-confidence interval. The plot shows that documents published between 2020 and 2022 have a significantly higher prevalence of Topic 1, "Freemium," and Topic 4, "Online Communities," than documents that were published between 1990 and 1999. Moreover, documents that were published between 2020 and 2022 have a significantly lower prevalence of Topic 6, "Branding." The other topics do not show significant differences. This means that, for Topics 2, 3, 5, 7, 8, and 9, the topic proportions do not differ according to the publication year in the document. Looking back at the grouping of the dendrogram in Section 4.1.2,

Table 5: Structural Topic Model Coefficients

	Topic								
	1	2	3	4	5	6	7	8	9
(Intercept)	0.047	0.058	0.082	0.176 *	0.202 .	0.169 .	0.064	0.051	0.150 .
	(0.084)	(0.081)	(0.074)	(0.086)	(0.111)	(0.094)	(0.082)	(0.088)	(0.091)
1990-1999	−0.007	0.137	−0.074	−0.168	−0.132	0.235 *	0.014	0.037	−0.038
	(0.100)	(0.108)	(0.090)	(0.103)	(0.127)	(0.105)	(0.097)	(0.112)	(0.108)
2000-2009	−0.015	0.059	0.026	−0.112	−0.054	0.031	−0.010	0.098	−0.020
	(0.092)	(0.088)	(0.083)	(0.094)	(0.117)	(0.102)	(0.089)	(0.100)	(0.100)
2010-2019	0.103	0.060	0.009	−0.051	−0.098	−0.120	0.053	0.067	−0.023
	(0.086)	(0.086)	(0.076)	(0.088)	(0.111)	(0.097)	(0.083)	(0.091)	(0.092)
2020-2022	0.108	0.041	0.013	−0.058	−0.082	−0.126	0.0230	0.114	−0.032
	(0.091)	(0.089)	(0.080)	(0.092)	(0.119)	(0.100)	(0.089)	(0.096)	(0.100)

Note: Signif. codes: '*' 0.05 '.' 0.1

Topic 1 and Topic 4 are situated on the "Digital" side of the dendrogram, while Topic 6 is not. This result suggests that there is an ongoing shift in the literature on free products from physical to digital.

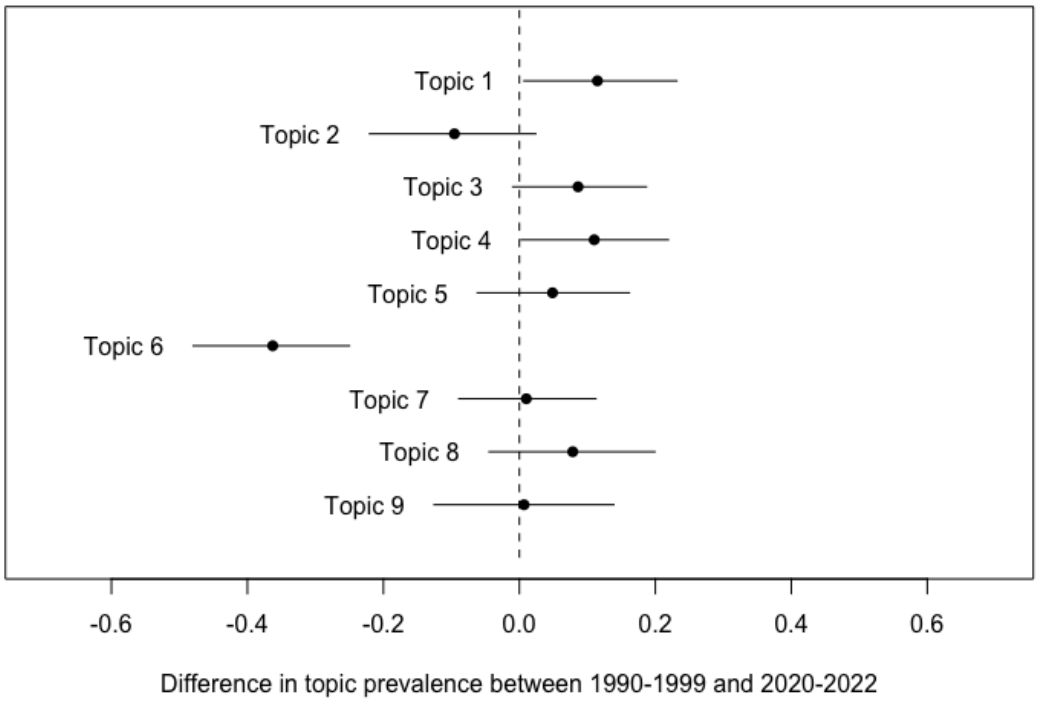


Figure 11: Structural Topic Model Time Effects

In addition to the contrast analysis, we can examine the topics independently over time. Figure 12 shows how the topic prevalence shifts over time per topic. The results are significant

at a level of 5%. Topics 1, 3, 4, 7, 8, and 9 have an increasing topic prevalence over the years. In the case of Topics 3 ("Online Consumer Behavior"), 8 ("Internet Companies"), and 9 ("Retail Promotions"), the increasing trend started in the interval of 2000-2009. This implies that these topics' expected topic proportions are increasing over time. The prevalence of Topics 1 ("Freemium Strategy") and 7 ("Zero Price") has increased more recently since the time interval 2010-2019. We find a decrease in topic prevalence in Topic 6 ("Branding") and a smaller decrease in Topic 2 ("Reward Programs"). In the case of Topics 4 ("Online Communities") and 5 ("Free Samples and Trials"), we measure the largest topic proportion in the time interval "<1989", followed by a big decline in "1990-1999". The proportions in Topic 4 and Topic 5 started rising again from "2000-2009" and "2010-2019", respectively. As discussed before, we put large effects in the interval "<1990" in perspective, given the relatively small number of articles. Summarized, we observe increasing trends in the literature on free products regarding Freemium Strategies, Online Consumer Behavior, Zero Price, Internet Companies, and Retail Promotions. On the other way around, we detect a decreasing trend concerning Reward Programs and Branding.

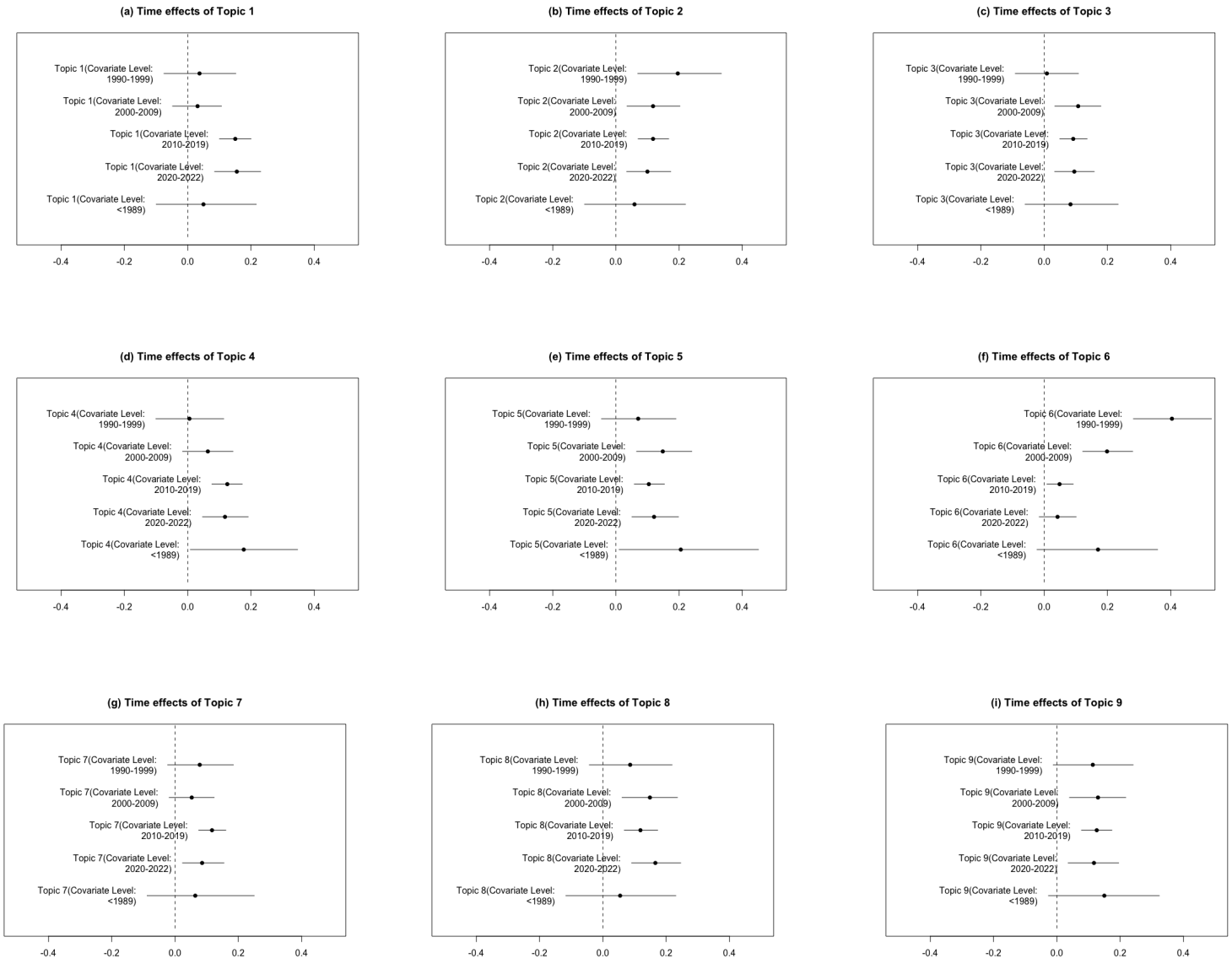


Figure 12: Time effects of Topic 1 - Topic 9

5 Limitations and further research

We must point out several limitations when we interpret the results of this study. First, a topic model is an exploratory tool to examine a corpus's underlying structure and does not offer one right solution. We chose a topic model based on several metrics, but we could have chosen another topic model if we had put more weight on another metric than semantic coherence.

Additionally, the topic model in this study was based only on the abstracts of the articles and did not consider the full content of the papers. A topic model that includes the full text of the articles may provide a more comprehensive understanding of the structure of the literature in this field.

It is also worth noting that the keywords included in the Scopus query can significantly influence the model. Changing the number or selection of keywords would result in a different set of articles and, therefore, a different outcome from the topic model. The same holds for including specific literature areas and types in the query.

While running the structural topic model is a fast and easy process, a significant amount of manual work is required before the analysis can be conducted. For example, in this study, we had to manually retrieve the abstracts, create a dataset with the complete list of authors for each publication, and conduct a manual review of each article to determine its relevance to the scope of the study. This suggests the potential for improving the efficiency and speed of an automatic literature review, particularly in the pre-processing stage.

Regarding the graph model, further research can investigate the benefits of including a full graph for an exploratory literature review of the metadata. Extra nodes such as articles, journals, and universities can be created. If the graph database contains more elements, researchers can use community and centrality algorithms to visualize the corpus' metadata better. We can improve our understanding of the author networks by linking authors to specific topics to see which authors have the strongest connections to certain topics. Additionally, we can enhance the citation analysis component of the networks by selecting the most highly cited articles within specific time intervals, rather than using a single citation threshold across all periods.

Lastly, this overview only consists of literature published between 1976 and July 2022. Given the increasing trend in digital products in the literature, it is encouraging to further investigate the literature and how it evolves. An algorithm that automatically renews the topic model after a specific time would be advantageous.

6 Conclusion

In this exploratory literature review, we aimed to identify the underlying structure and evolution of the existing research on free products. To do so, we developed a topic model based on the abstracts of 279 articles in this field.

Our metadata analysis revealed that this field is widely studied across various author net-

works and journals. We compared Latent Dirichlet Allocation (LDA) topic models and Structural Topic Models (STMs) to determine which approach best fits our research question. We found that STMs were more suitable because they allow for the correlation between topics and the inclusion of covariates.

We created eight pre-processing setups, each with a different cleaning process and covariate, and fit 28 models for each setup to determine the optimal number of topics. After evaluating the models, we selected a model with nine topics, a time interval covariate, and no stemming or stop word removal.

We labeled the topics by combining the insights from the highest FREX terms and the most probable documents. Most labels were easy to identify based on the clear semantic meaning of the terms. However, some labels, such as "Zero Pricing" (Topic 7) and "Internet Companies" (Topic 8), were more challenging to group into a single cluster.

We used a dendrogram to cluster the topics based on their semantic coherence and observed two general trends in the field. On the one hand, we saw a range of marketing techniques that utilize free products. On the other hand, we observed topics primarily situated in a digital environment.

When we analyzed the temporal covariate, the topic model indicated an evolution over time. We observed a significant difference between the time intervals "1990-1999" and "2020-2022" for three topics: "Freemium Strategy" (Topic 1) and "Online Consumer Behavior" (Topic 3) are significantly more prevalent in the latter period, while "Branding" (Topic 6) was more common in the earlier period. This result suggests a trend toward a greater focus on free products in digital environments. When examining the individual time effects of the topics, we also found that some topics have remained prevalent throughout all time intervals, such as "Reward Programs" (Topic 2), "Free Samples and Trials" (Topic 5), and "Retail Promotions" (Topic 9).

Our results indicate that using free products as a marketing tool has a long history and continues to be a common practice. In recent years, however, we have seen an increase in companies using free products as a business strategy, particularly online businesses. This shift reflects the growing importance of free products in today's business landscape.

References

Appel, G., Libai, B., Muller, E., and Shachar, R. (2020). On the monetization of mobile apps. *International Journal of research in Marketing*, 37(1):93–107.

Arora, S., Ter Hofstede, F., and Mahajan, V. (2017). The implications of offering free versions for the performance of paid mobile apps. *Journal of Marketing*, 81(6):62–78.

Asmussen, C. B. and Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1):1–18.

Asmussen, C. B. and Møller, C. (2020). Enabling supply chain analytics for enterprise information systems: a topic modelling literature review and future research agenda. *Enterprise Information Systems*, 14(5):563–610.

Bai, X., Zhang, X., Li, K. X., Zhou, Y., and Yuen, K. F. (2021). Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*, 102:11–24.

Bapna, R. and Umyarov, A. (2015). Do your online friends make you pay? a randomized field experiment on peer influence in online social networks. *Management Science*, 61(8):1902–1920.

Bhargava, H. K. and Chen, R. R. (2012). The benefit of information asymmetry: When to sell to informed customers? *Decision Support Systems*, 53(2):345–356.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bründl, S., Matt, C., Hess, T., and Engert, S. (2022). How synchronous participation affects the willingness to subscribe to social live streaming services: The role of co-interactive behavior on twitch. *European journal of information systems*, pages 1–18.

Buratto, A., Cesaretto, R., and De Giovanni, P. (2019). Consignment contracts with cooperative programs and price discount mechanisms in a dynamic supply chain. *International Journal of Production Economics*, 218:72–82.

Chen, H., Duan, W., and Zhou, W. (2017). The interplay between free sampling and word of mouth in the online software market. *Decision Support Systems*, 95:82–90.

Chen, J., Fan, M., and Li, M. (2016). Advertising versus brokerage model for online trading platforms. *Mis Quarterly*, 40(3):575–596.

Chen, L., Marsden, J. R., and Zhang, Z. (2013). Reliability (or "lack thereof") of on-line preference revelation: a controlled experimental analysis. *Decision support systems*, 56:270–274.

Dey, D., Lahiri, A., and Liu, D. (2013). Consumer learning and time-locked trials of software products. *Journal of Management Information Systems*, 30(2):239–268.

Dickinson, J. R. and Wilby, C. P. (1997). Concept testing with and without product trial. *Journal of Product Innovation Management: AN INTERNATIONAL PUBLICATION OF THE PRODUCT DEVELOPMENT & MANAGEMENT ASSOCIATION*, 14(2):117–125.

Ding, Y. and Zhang, Y. (2020). Hiding gifts behind the veil of vouchers: on the effect of gift vouchers versus direct gifts in conditional promotions. *Journal of Marketing Research*, 57(4):739–754.

Dubé, J.-P., Luo, X., and Fang, Z. (2017). Self-signaling and prosocial behavior: A cause marketing experiment. *Marketing Science*, 36(2):161–186.

Erdem, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing science*, 15(4):359–378.

Foster, D. (2019). *Generative deep learning: teaching machines to paint, write, compose, and play*. O'Reilly Media.

Gabel, S. and Guhl, D. (2022). Comparing the effectiveness of rewards and individually targeted coupons in loyalty programs. *Journal of Retailing*, 98(3):395–411.

Gedenk, K. and Neslin, S. A. (1999). The role of retail promotion in determining future brand loyalty: Its effect on purchase event feedback. *Journal of Retailing*, 75(4):433–459.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Gu, X., Kannan, P., and Ma, L. (2018). Selling the premium in freemium. *Journal of Marketing*, 82(6):10–27.

Gupta, K., Su, Y., Kunkel, T., and Funk, D. (2022). The super engagers of freemium gamified services: using multimethod approach to examine why highly interactive consumers become paying consumers. *Internet Research*, (ahead-of-print).

Hamari, J., Alha, K., Järvelä, S., Kivikangas, J. M., Koivisto, J., and Paavilainen, J. (2017). Why do players buy in-game content? an empirical study on concrete purchase motivations. *Computers in Human Behavior*, 68:538–546.

Hamari, J., Malik, A., Koski, J., and Johri, A. (2019). Uses and gratifications of pokémon go: why do people play mobile location-based augmented reality games? *International Journal of Human–Computer Interaction*, 35(9):804–819.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer.

Hsu, C.-L., Lin, J. C.-C., and Chiang, H.-S. (2013). The effects of blogger recommendations on customers' online shopping intentions. *Internet Research*.

Jain, D., Mahajan, V., and Muller, E. (1995). An approach for determining optimal product sampling for the diffusion of a new product. *Journal of Product Innovation Management*, 12(2):124–135.

Karmakar, B., Liu, P., Mukherjee, G., Che, H., and Dutta, S. (2022). Improved retention analysis in freemium role-playing games by jointly modelling players' motivation, progression and churn. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Kempf, D. S. (1999). Attitude formation from product trial: Distinct roles of cognition and affect for hedonic and functional products. *Psychology & Marketing*, 16(1):35–50.

Kempf, D. S. and Smith, R. E. (1998). Consumer processing of product trial and the influence of prior advertising: A structural modeling approach. *Journal of Marketing Research*, 35(3):325–338.

Kim, B.-D., Shi, M., and Srinivasan, K. (2001). Reward programs and tacit collusion. *Marketing Science*, 20(2):99–120.

Kourandi, F., Krämer, J., and Valletti, T. (2015). Net neutrality, exclusivity contracts, and internet fragmentation. *Information Systems Research*, 26(2):320–338.

Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, 87:105–122.

Kumar, S., Jacob, V. S., and Sriskandarajah, C. (2006). Scheduling advertisements on a web page to maximize revenue. *European journal of operational research*, 173(3):1067–1089.

Lafferty, J. and Blei, D. (2005). Correlated topic models. *Advances in neural information processing systems*, 18.

Lammers, H. B. (1991). The effect of free samples on immediate consumer purchase. *Journal of Consumer Marketing*.

Lavoie, R. and Main, K. (2022). Optimizing product trials by eliciting flow states: the enabling roles of curiosity, openness and information valence. *European Journal of Marketing*.

Lee, S. and Yi, Y. (2017). "seize the deal, or return it losing your free gift": The effect of a gift-with-purchase promotion on product return intention. *Psychology & Marketing*, 34(3):249–263.

Lee, S., Zhang, J., and Wedel, M. (2021). Managing the versioning decision over an app's lifetime. *Journal of Marketing*, 85(6):44–62.

Lindstedt, N. C. (2019). Structural topic modeling for social scientists: A brief case study with social movement studies literature, 2005–2017. *Social Currents*, 6(4):307–318.

Liu, Z., Lin, Z., Zhang, Y., and Tan, Y. (2022). The signaling effect of sampling size in physical goods sampling via online channels. *Production and Operations Management*, 31(2):529–546.

Mehra, A. and Saha, R. L. (2018). Utilizing public betas and free trials to launch a software product. *Production and Operations Management*, 27(11):2025–2037.

Meng, Z., Hao, L., and Tan, Y. (2021). Freemium pricing in digital games with virtual currency. *Information Systems Research*, 32(2):481–496.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Nan, G., Wu, D., Li, M., and Tan, Y. (2018). Optimal freemium strategy for information goods in the presence of piracy. *Journal of the Association for Information Systems*, 19(4):3.

Nicolau, J. L. (2012). Battle royal: Zero-price effect vs relative vs referent thinking. *Marketing Letters*, 23(3):661–669.

Park, Y. and Yi, Y. (2019). When free gifts hurt the promoted product: The influence of product-gift fit on activating persuasion knowledge and devaluating the promoted product. *European Journal of Marketing*.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). Stm: An r package for structural topic models. *Journal of Statistical Software*, 91:1–40.

Roberts, M. E., Stewart, B. M., Tingley, D., and Airoldi, E. M. (2013). The structural topic model and applied social science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.*

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.

Roy, R., Chintagunta, P. K., and Haldar, S. (1996). A framework for investigating habits,"the hand of the past," and heterogeneity in dynamic brand choice. *Marketing science*, 15(3):280–299.

Schofield, A., Magnusson, M., and Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, short papers*, pages 432–436.

Schofield, A. and Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Shampanier, K., Mazar, N., and Ariely, D. (2007). Zero as a special price: The true value of free products. *Marketing science*, 26(6):742–757.

Shang, G., Ferguson, M. E., and Galbreth, M. R. (2019). Where should i focus my return reduction efforts? empirical guidance for retailers. *Decision Sciences*, 50(4):877–909.

Shen, C.-C. and Chiou, J.-S. (2009). The effect of community identification on attitude and intention toward a blogging community. *Internet Research*.

Sheng, L., Ryan, C. T., Nagarajan, M., Cheng, Y., and Tong, C. (2022). Incentivized actions in freemium games. *Manufacturing & Service Operations Management*, 24(1):275–284.

Spiekermann, S. and Korunovska, J. (2017). Towards a value theory for personal data. *Journal of Information Technology*, 32(1):62–84.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In *Handbook of latent semantic analysis*, pages 439–460. Psychology Press.

Taddy, M. (2012). On estimation and selection for topic models. In *Artificial Intelligence and Statistics*, pages 1184–1193. PMLR.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112.

Wang, S. and Özkan-Seely, G. F. (2018). Signaling product quality through a trial period. *Operations Research*, 66(2):301–312.

Washburn, J. H., Till, B. D., and Priluck, R. (2004). Brand alliance and customer-based brand-equity effects. *Psychology & Marketing*, 21(7):487–508.

Zhou, W. and Duan, W. (2015). An empirical study of how third-party websites influence the feedback mechanism between online word-of-mouth and retail sales. *Decision Support Systems*, 76:14–23.

Zhu, D. H., Chang, Y. P., and Chang, A. (2015). Effects of free gifts with purchase on online purchase satisfaction: The moderating role of uncertainty. *Internet Research*.

A Author Collaboration Graphs

Figure 13: Author networks with more than one publication

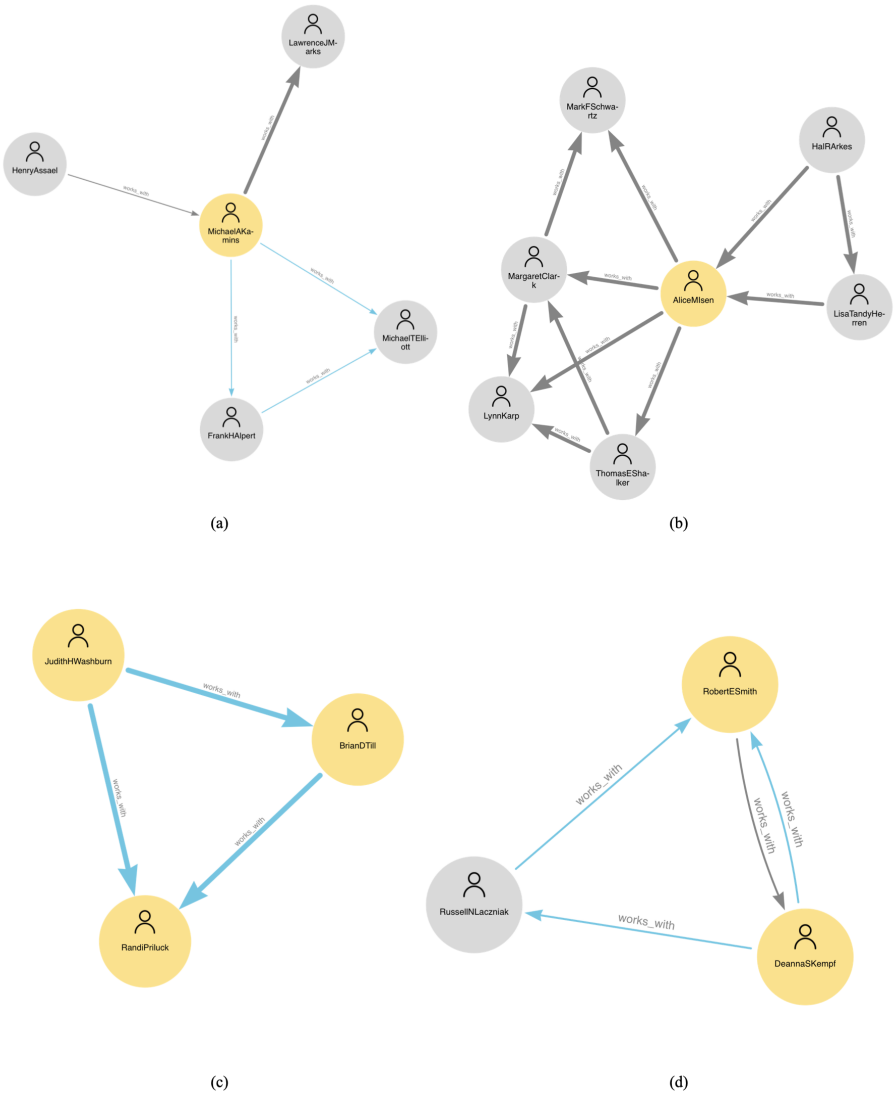


Figure 13: Author networks with more than one publication - continued

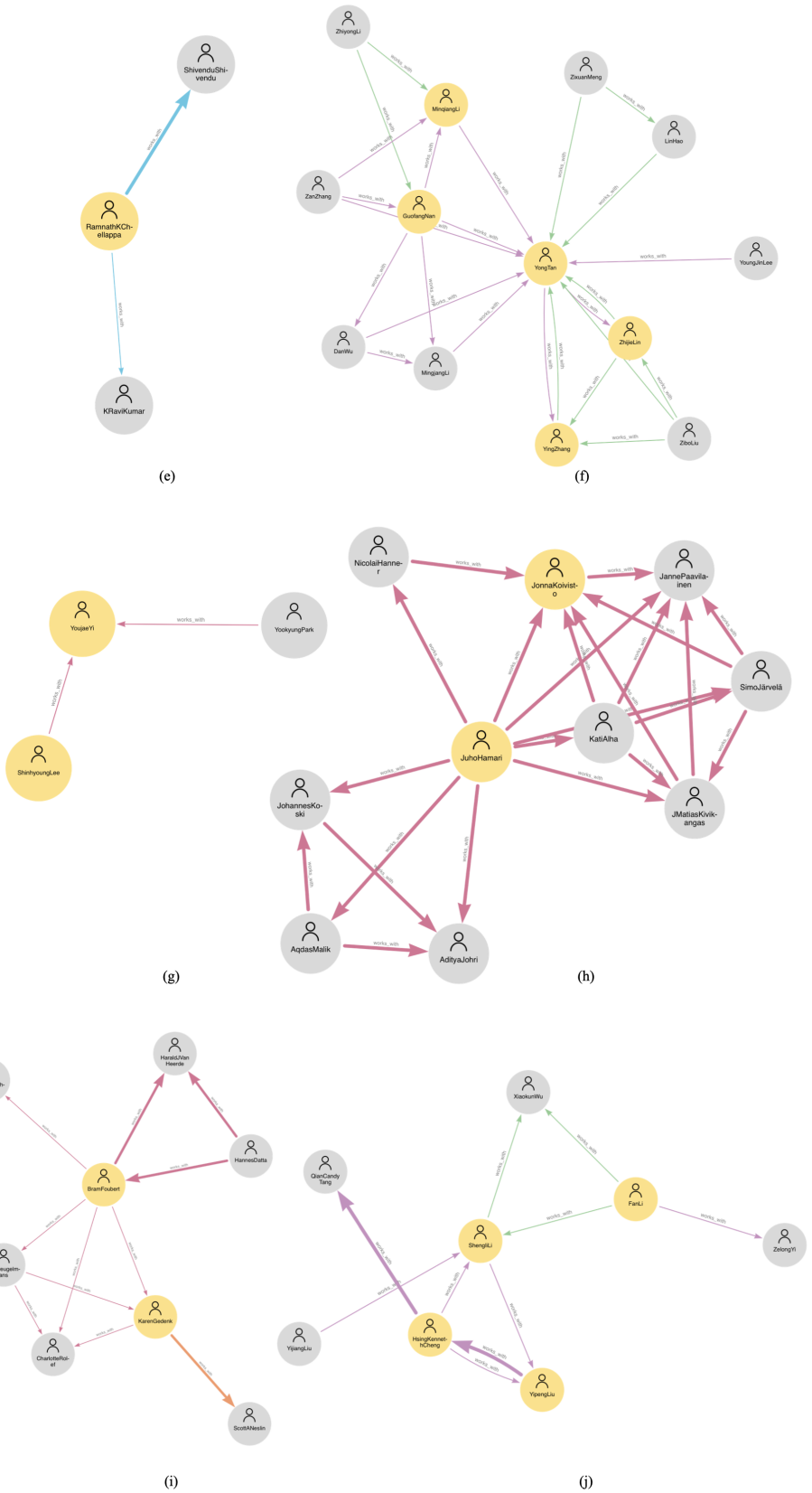
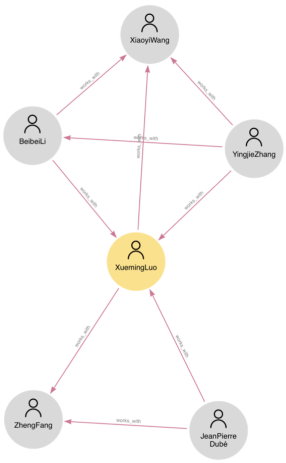
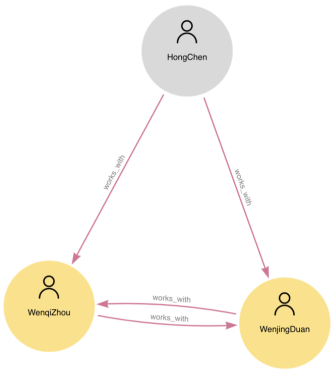


Figure 13: Author networks with more than one publication - continued

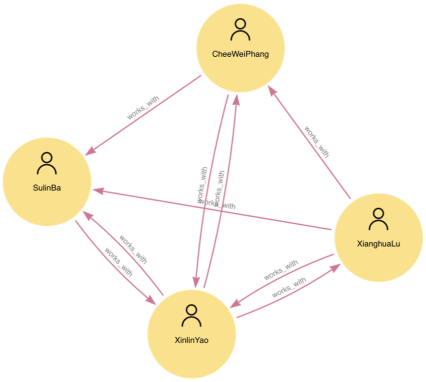
(k)



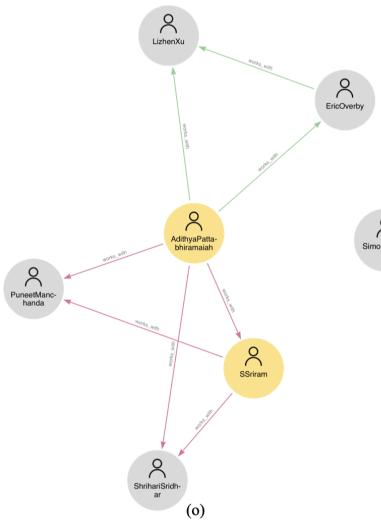
(l)



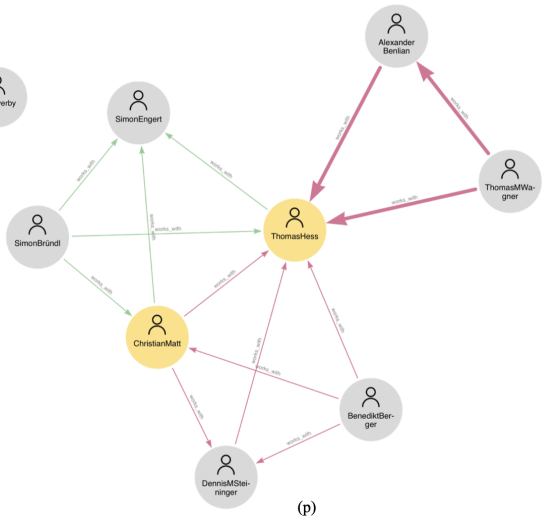
(m)



(n)

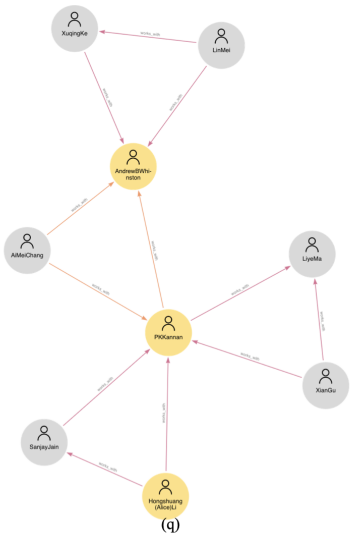


(o)

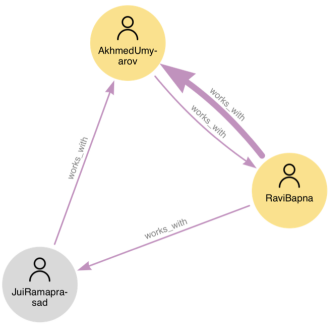


(p)

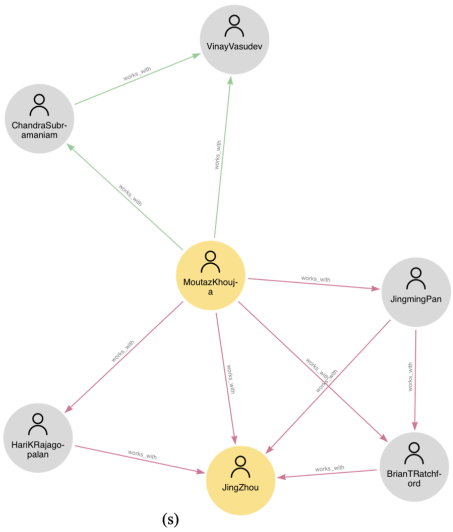
Figure 13: Author networks with more than one publication - continued



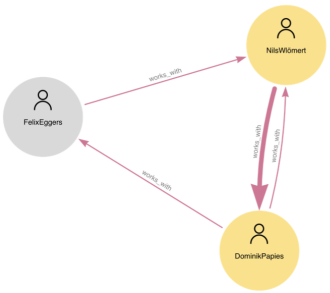
(q)



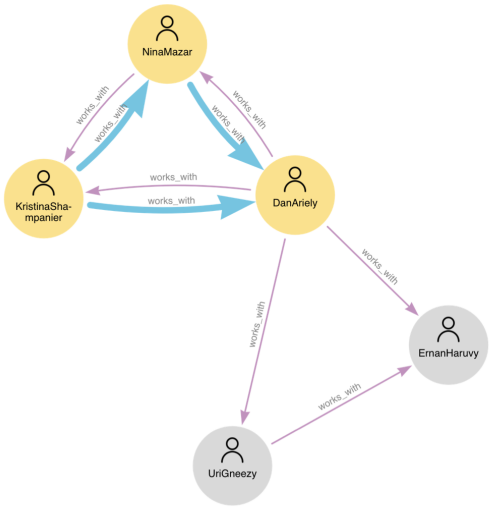
(r)



(s)



(t)



(u)

(v)

Figure 13: Author networks with more than one publication - continued

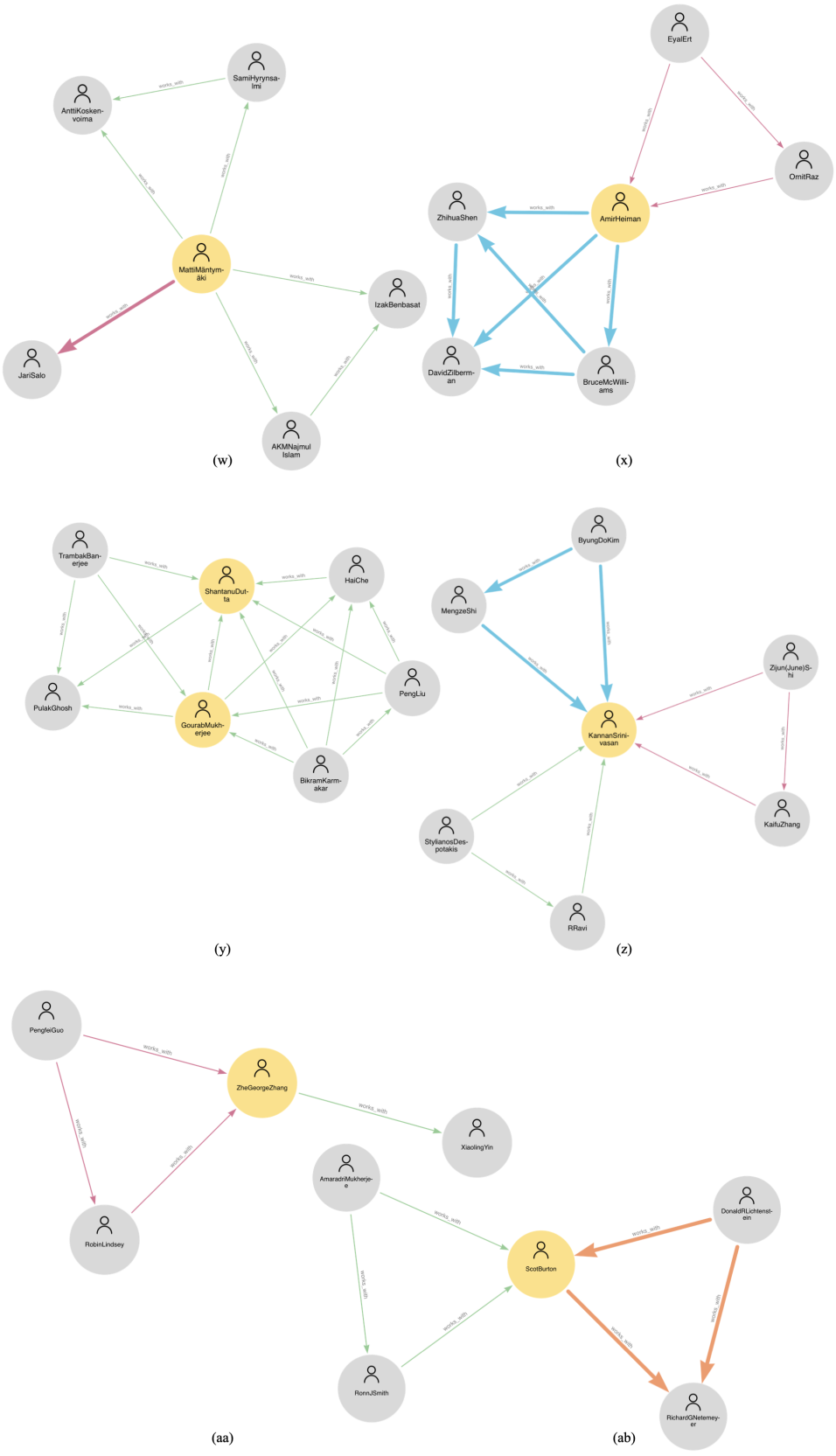
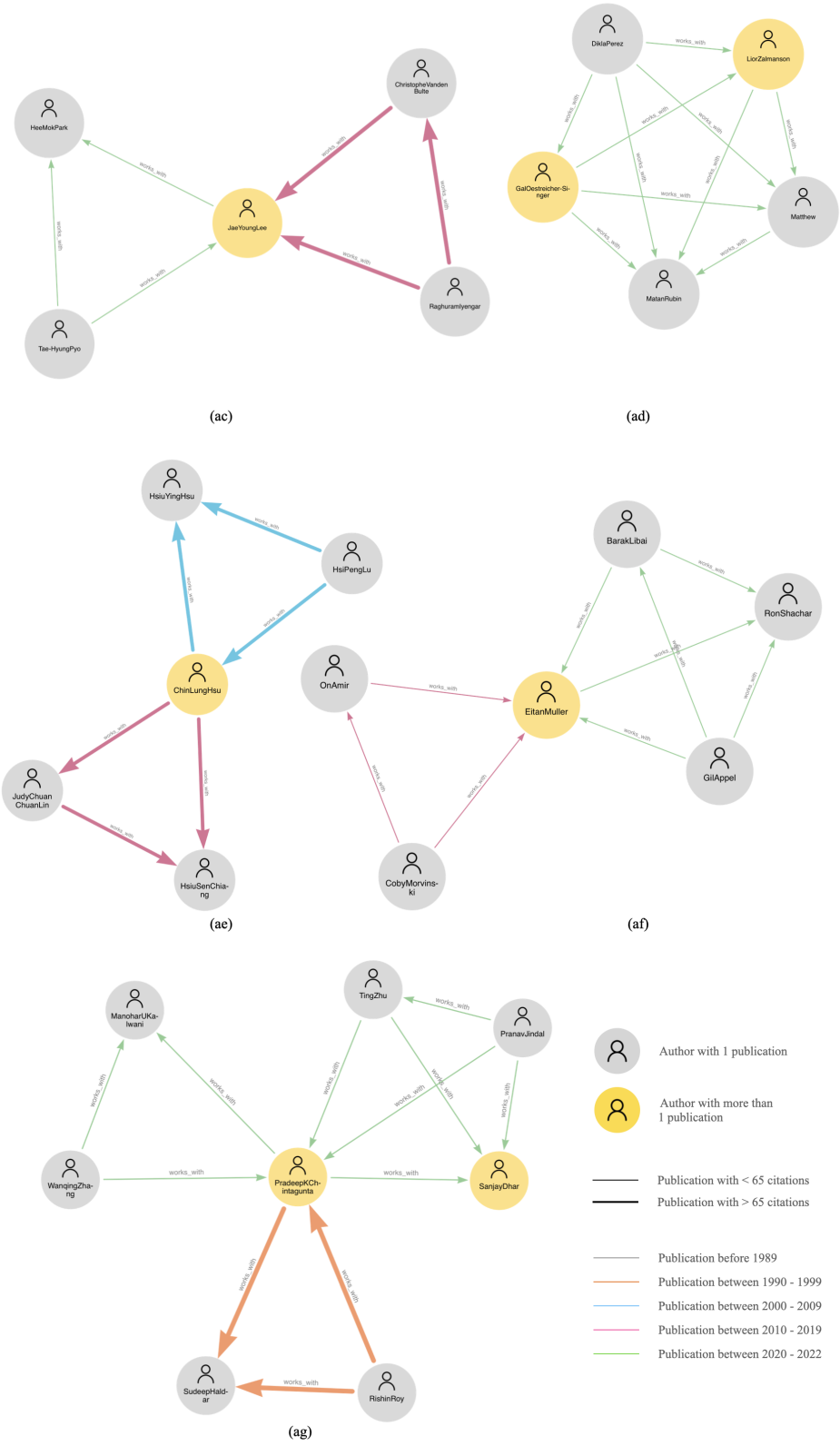


Figure 13: Author networks with more than one publication - continued



B Topic model evaluation plots

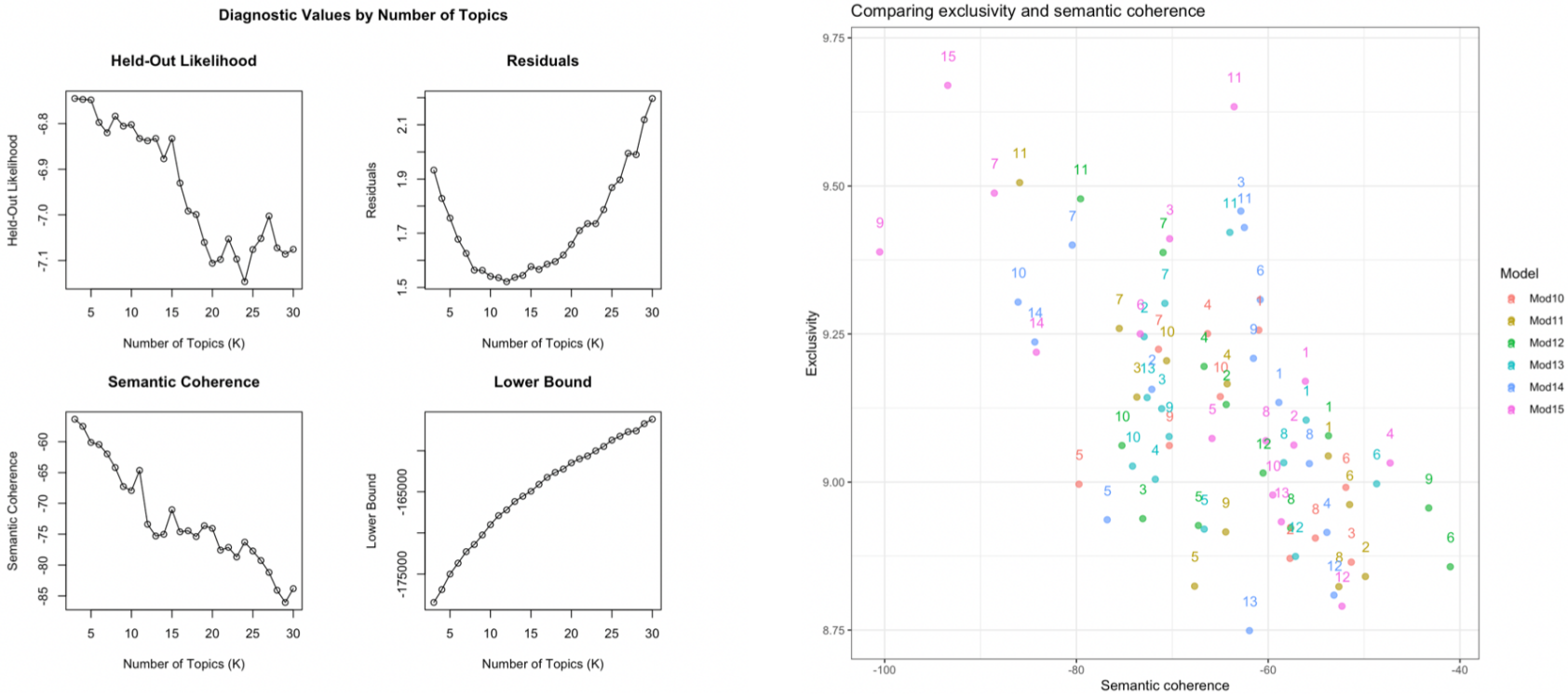


Figure 14: Topic Model: time interval covariate, with stemming and with removing stop words

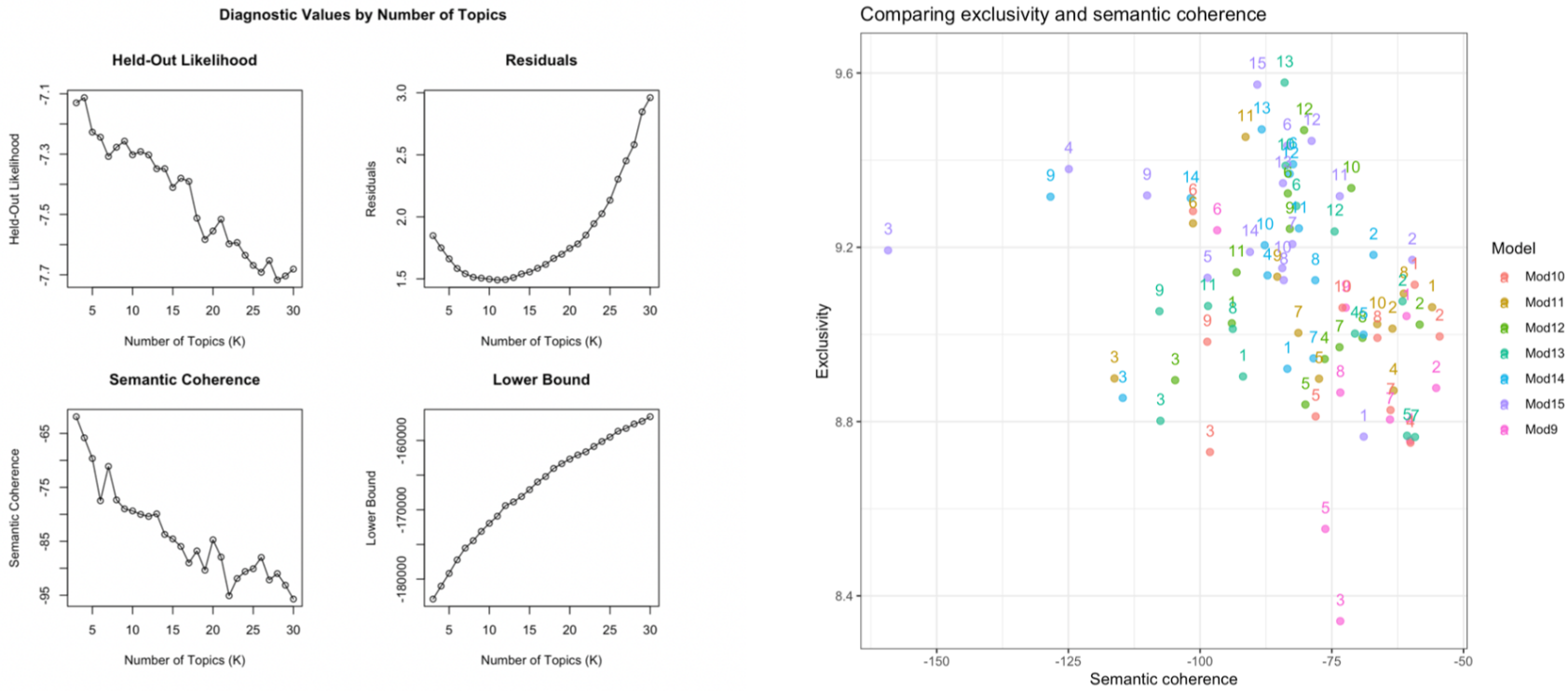


Figure 15: Topic Model: time interval covariate, without stemming, with stop word removal

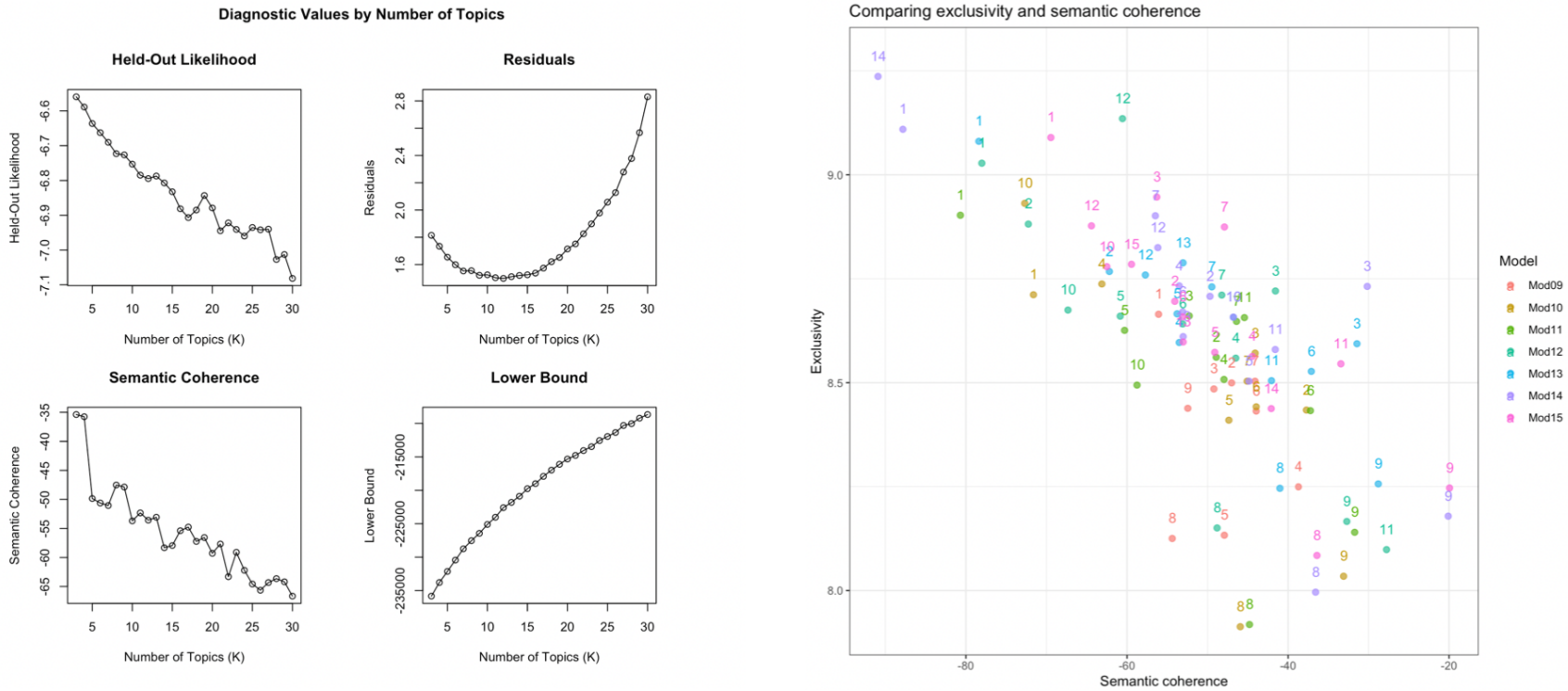


Figure 16: Topic Model: time interval covariate, no stemming, no stopwords removal

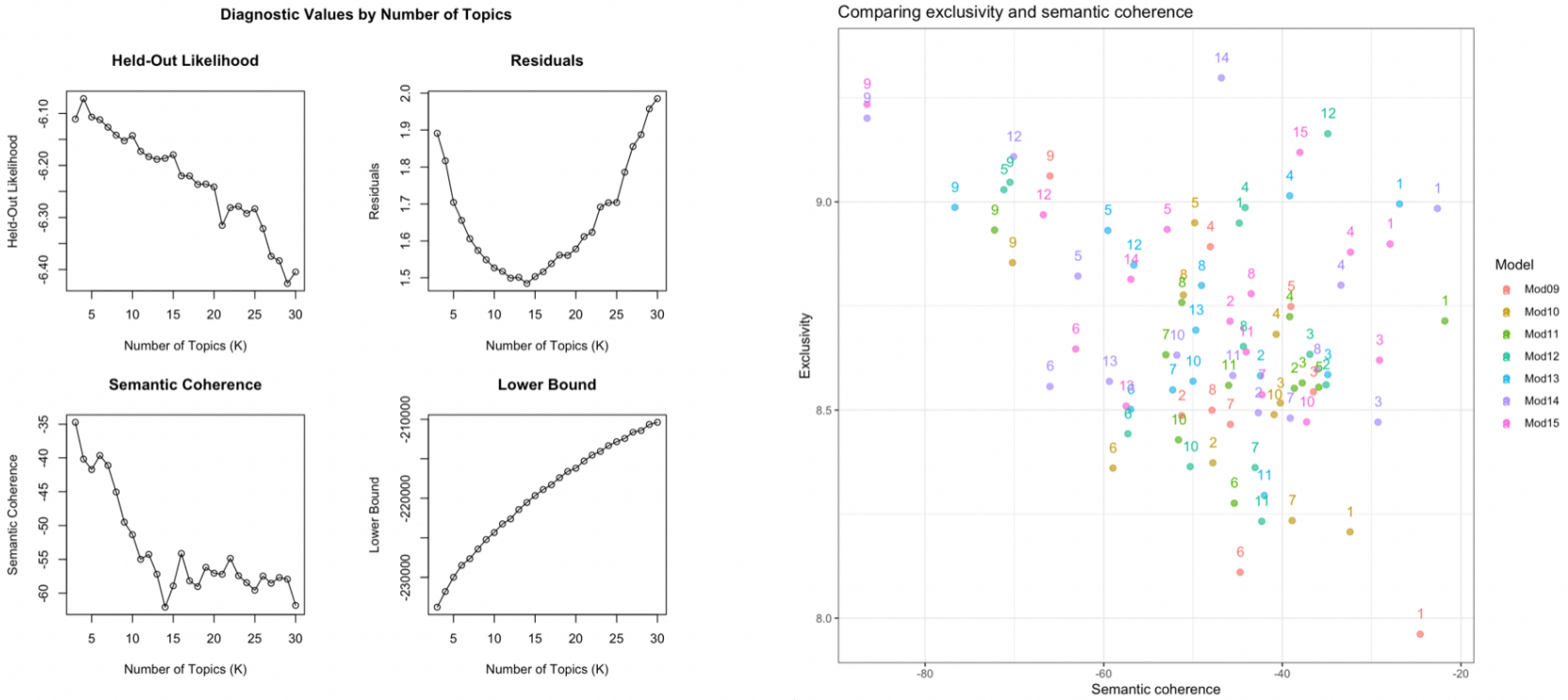


Figure 17: Topic Model: time interval covariate, with stemming, no stopwords removal

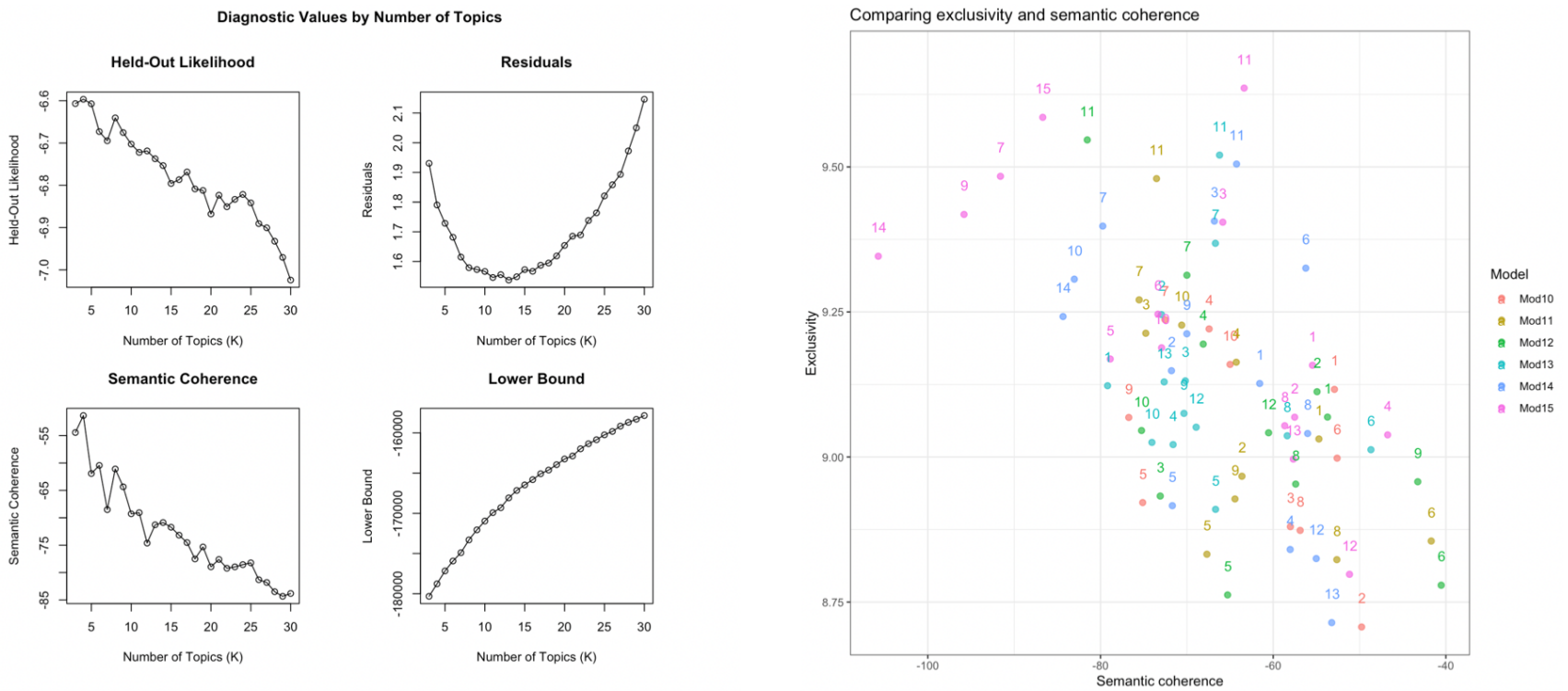


Figure 18: Topic Model: continuous time covariate, with stemming, stopword removal

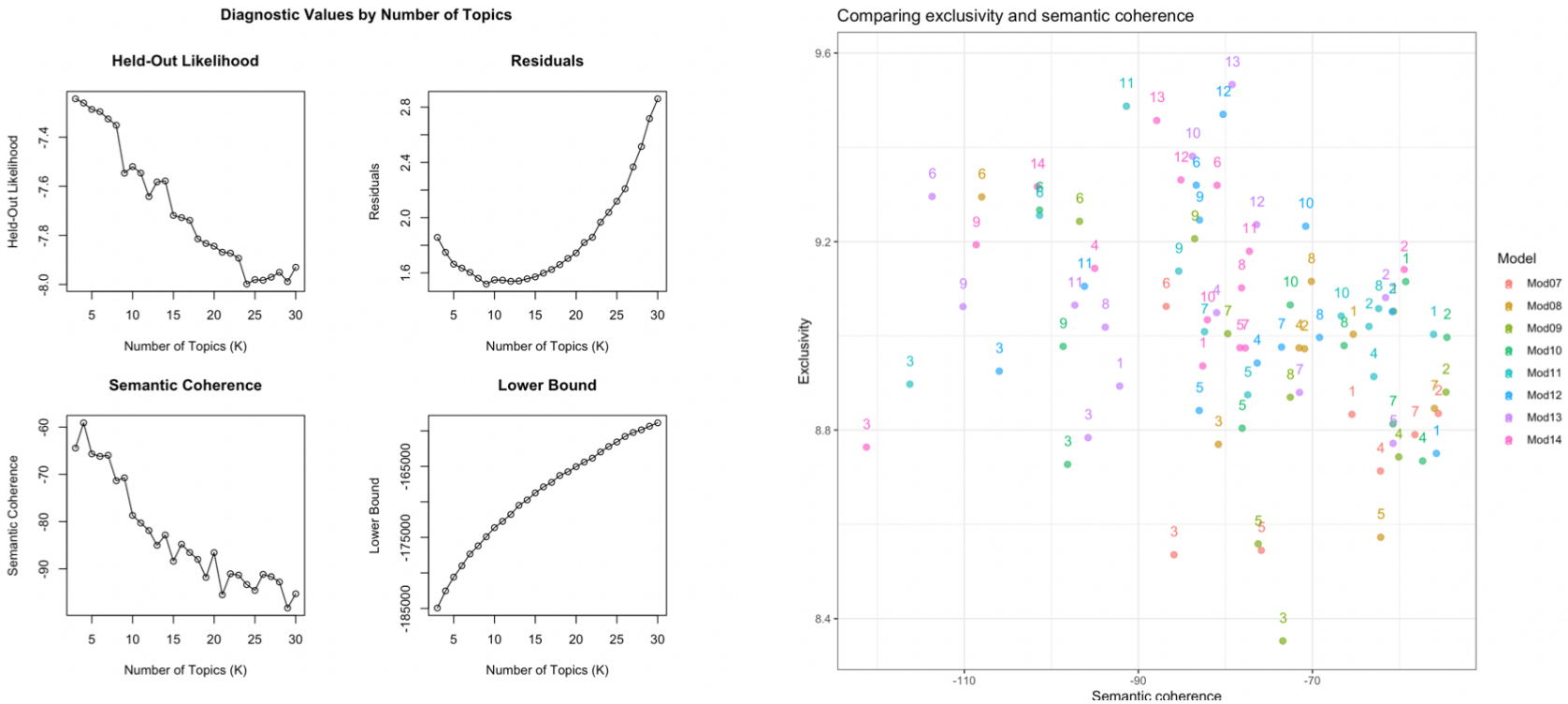


Figure 19: Topic Model: continuous time covariate, without stemming, with stopwords removal

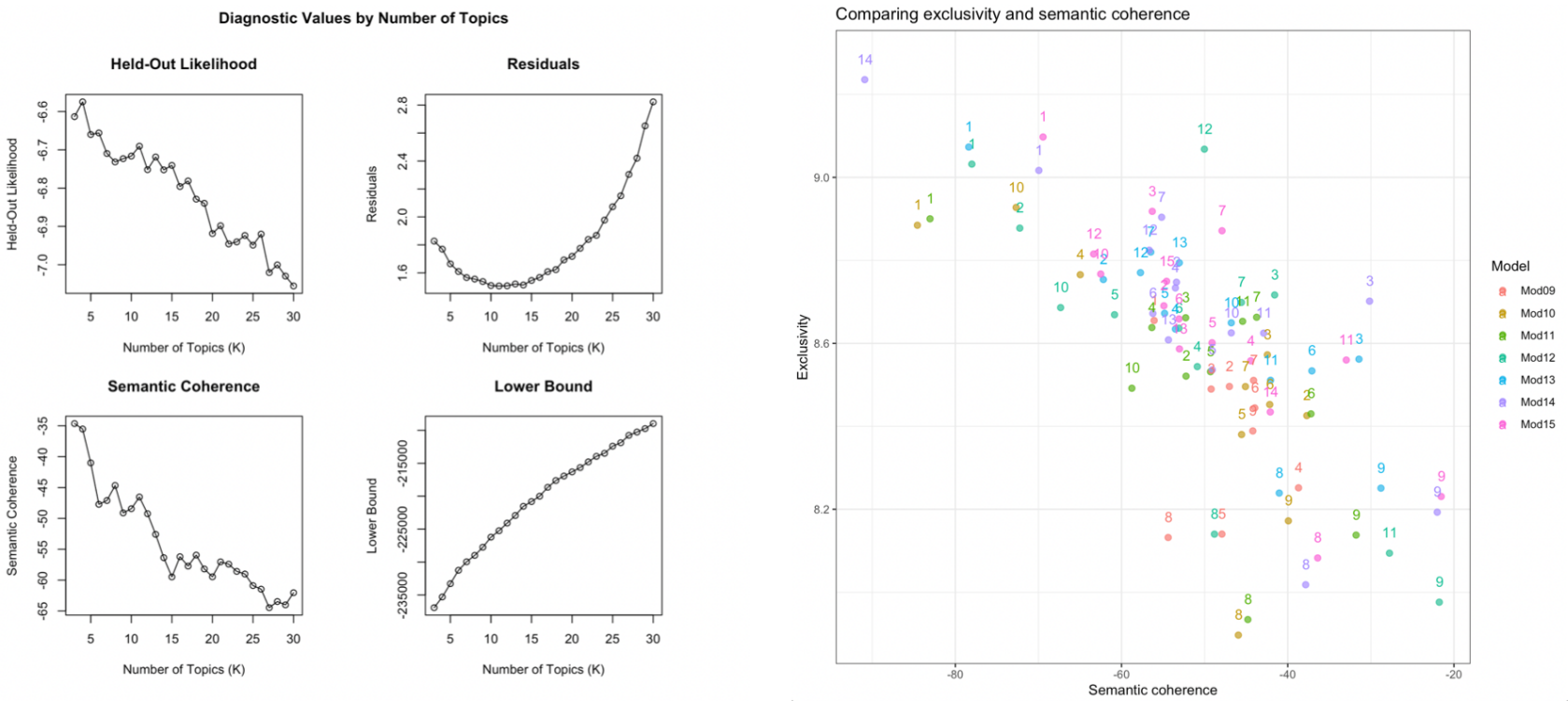


Figure 20: Topic Model: continuous time covariate, no stemming, no stopwords removal

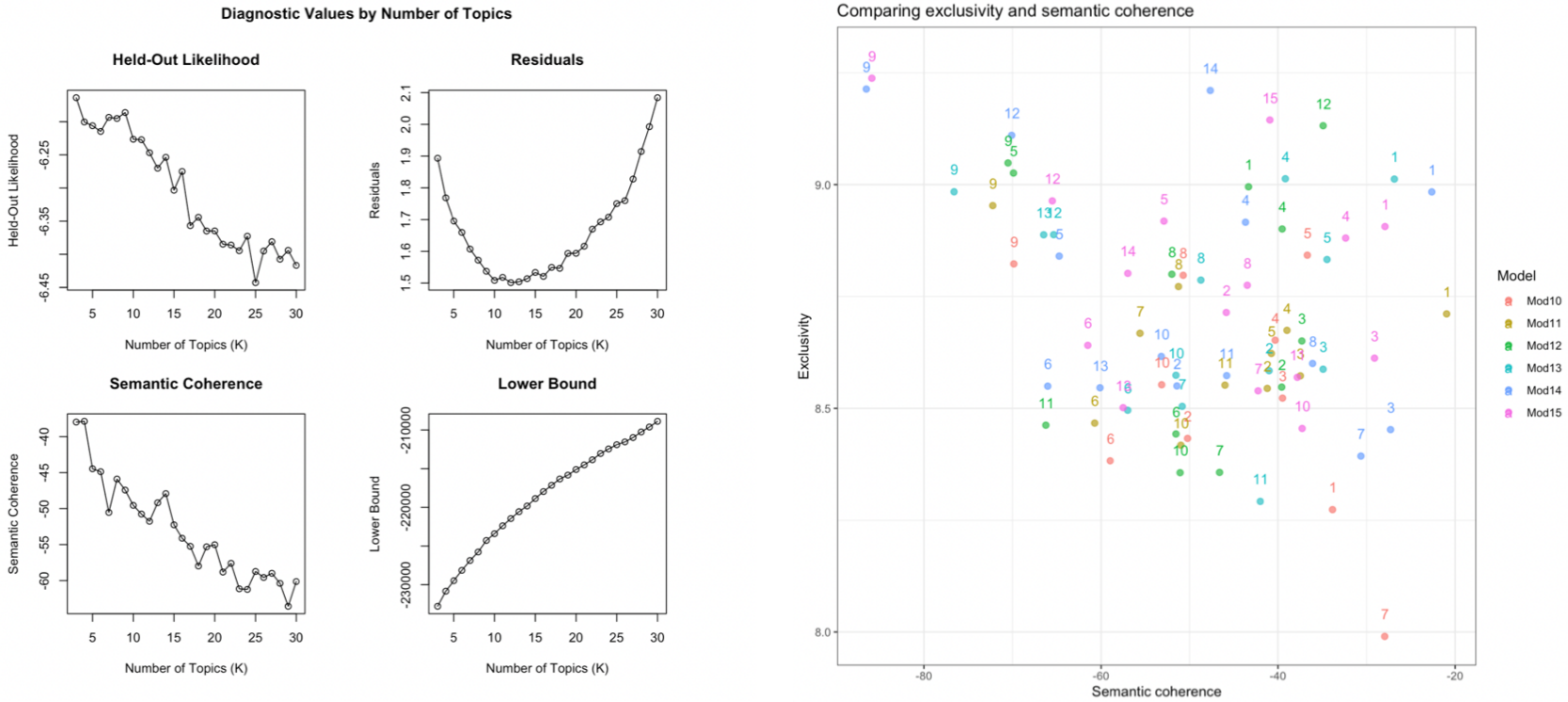


Figure 21: Topic Model: continuous time covariate, with stemming, no stopwords removal