Analysing the History of Autism Spectrum Disorder using Topic Models

Authors

***Abstract*—We describe a novel framework for the discovery of underlying topics of a longitudinal collection of scholarly data, and the tracking of their lifetime and popularity over time. Unlike the social media or news data, as the topic nuances in science result in new scientific directions to emerge, a new approach to model the longitudinal literature data is using topics which remain identifiable over the course of time. Current studies either disregard the time dimension or treat it as an exchangeable covariate when they fix the topics over time or do not share the topics over epochs when they model the time naturally. We address these issues by adopting a non-parametric Bayesian approach. We assume the data is partially exchangeable and divided it into consecutive epochs. Then, by fixing the topics in a recurrent Chinese restaurant franchise, we impose a static topical structure on the corpus such that the they are shared across epochs and the documents within epochs. We demonstrate the effectiveness of the proposed framework on a collection of medical literature related to autism spectrum disorder. We collect a large corpus of publications and carefully examining two important research issues of the domain as case studies. Moreover, we make the results of our experiment and the source code of the model, freely available to aid other researchers by analysing the results or applying the model to their data collections.**

***Index Terms*—Bayesian nonparametrics, data mining, autism spectrum disorder**

I. INTRODUCTION

What are the topics of this domain? Is this topic still *hot*? – Questions that all researchers have asked themselves at least once in their professional life. While the answer to such questions might be apparent to experienced researchers, those less senior or from cross-disciplinary studies are not usually aware of the entire research space of their domain. Even experienced scientists often have a high-level view of their domain topics and apart from their niche of expertise, they are not entirely aware of other fine-grained topics that span their research field; particularly in the biomedical domain, where the most cutting-edge science occurs in multidisciplinary studies and the knowledge is continuously growing. These reasons stress why understanding the underlying topical structure of science has always been of interest to researchers.

Topic models offer a statistical framework which along with the availability of digital archives of scholarly literature, solves this problem elegantly. They closely conform to the process of our writing. To compose an article, we usually choose a handful of topics and then write about those topics. We do the same thing in a probabilistic framework in topic modelling. We associate a probability distribution over words with each topic, and a probability distribution over topics with each document. Then, each document is constructed by repeatedly choosing a topic from the document-topic distribution, and drawing a word from topic-word distribution.

Since topic models enable us to learn these hidden distributions from large archives of documents, they are useful tools to organize, search, and understand the vast amount of information available from research repositories such as PubMed or Scopus. They provide a quantitative map of the key topics in a research field and the papers that contributed to those mostly, by analysing the *entire* research field; something desirable for every researcher but difficult to achieve individually.

Topic models have been applied in some domains such as digital humanities, linguistics, and cognition for this purpose and proved to be useful [10], [11], [7]. Along the same lines, this work explores the underlying topics of the autism spectrum disorder (ASD) and investigates how they have changed over the last four decades. Although ASD is a relatively new research domain, there have been a few shifts in our understanding of its nature, symptoms, contributing factors, and its treatments which have resulted in a broad and extensive body of research with sometimes hypotheses with little evidence, even in direct contradiction.

More specifically, we describe a general framework for the analysis of research literature capable of (i) discovering the underlying topical structure, (ii) tracking the lifetime of topics over time, and (iii) tracking the popularity of topics over time. The proposed framework uses Recurrent Chinese Restaurant Franchise [**?**] with the topics that are fixed over epochs in order to maintain identifyability over time. fixed topics. The effectiveness of our approach is demonstrated on a large longitudinal data corpus of 46,218 articles published over the course of four decades. In addition to afro-mentioned technical contributions, our further contributions are (a) a visualisation of the results, available to the ASD researchers to explore the topics and relevant papers in the corpus interactively[1] and (b) the source code of the model which is freely available to other researchers[2].

II. PREVIOUS RELATED WORK

As mentioned earlier, others have used topic modelling to uncover the thematic structure of a research domain. Here, we review the most notable relevant previous studies.

Hall et al. used topic models to analyse the themes and their popularity over time in the field of computational linguistics [7]. Their approach first obtains the topics of a corpus of published papers by fitting a Latent Dirichlet Allocation (LDA). Then a simple post-processing is applied to

[1] http://google.com - link removed due to the blind review process

[2] http://google.com - link removed due to the blind review process

find the popularity of each topic over time. The results show which topics have received more research interest over time, as well as those with fewer papers associated with them, hence less interest. For example, they showed that in the context of computational linguistics, *probabilistic modelling* has steadily attracted more interest from 1988. This framework has soon become the standard approach for investigating the rise and fall of ideas in different fields of science. For example, it was used on a corpus of humanities classics of the last century [10] or the published research on cognition over the last 40 years [11]. However, this model suffers from two limitations: using a parametric topic model and disregarding the temporal order of the publications.

LDA is a parametric model i.e. we need to define its complexity by specifying the number of topics. Thus, obtaining the set of topics that describes the data best, needs fitting multiple models and then model selection. It is a computationally expensive procedure. Hierarchical Dirichlet Process (HDP) [15] addresses this limitation by using a non-parametric prior on the parameter space of the mixture components. This way the data dictates the complexity of the model and the right number of topics is automatically inferred.

One can simply replace the LDA in Hall et al. framework with HDP to infer the right number of topics. But there is yet another limitation that needs to be addressed. In Hall et al. approach, the temporal order of the data is disregarded. In other words, our prior belief on the distribution of topics is the same for two papers 50 years apart. This is not an approopriate assumption as new topics emerge, the popularity of topics change, and topics disappear over time.

There are two ways to address this limitation. The first is discretizing the time into multiple epochs and fitting a topic model to each epoch [4], [17], [12], [19], [1]. Usually, a Markovian assumption is made to relate the parameters of adjacent topic models and governs their changes. As the result, those models of this group that are used in the context of literature analysis opted to use similarity measures between topics of adjacent epochs to be able to follow a topic over time [1], [9]. This way, they can model the birth and death of topics, as well as their evolution. The proposed approach in [1] can model the split and merge behaviour of topics too.

The second method is modelling the time jointly with the data as in [5], [18], [17]. They treat the timestamp of observations as random variables and infer a distribution over time for each topic. For example Wang et al. proposed Topics Over Time (TOT), assuming the timestamps are drawn from Beta distributions [18] . This model has two significant drawbacks. Since it is an extension of LDA, it is parametric and because of Beta priors, it limits the type of behaviours that the popularity of a topic can show. Dubey et al. [5] extended TOT by placing non-parametric priors on topics and time stamps, allowing infinite mixtures of topics for documents, and infinite mixtures of time distributions for topics. Although both limitations are addressed in this model, the way they treat the data is not natural. They relax the order of the data by attaching the timestamps to the observations and assuming complete exchangeability of the tuples of data-timestamps. A more natural way to solve the problem is keeping the order of the data as proposed in the first group.

Our proposed method adopts the idea of time discretization similar to the first group but in contrast with them, the mixing components are shared across epochs. This way we avoid using similarity measures and thresholding to find the continuum of a topic over time. Therefore, unlike modelling longitudinal corpora in social media or news that topics are desired to evolve over time, we disregard the evolution. Because in scientific literature, changes in topics often result in new topics to emerge. This is explained further in Section III.

III. Methodology

We begin this section by reviewing the underlying theory of Recurrent Chinese Restaurant Franchise model (RCRF). Then we turn our attention to the main technical contribution of our work and explain how RCRF was used to discover the topical structure of a literature corpus.

A. Dirichlet process

The Dirichlet Process (DP) [6] is a distribution over a space of random probability measures. Therefore, its draws can be interpreted as random probability distributions. More formally, a Dirichlet Process $\mathrm{DP}(\gamma, H)$ is defined as the distribution of random probability measure G over a measurable space $(\Theta, \mathcal{B})$ such that for any finite measurable partition $(A_1, A_2, \dots, A_r)$ of Θ, the random vector $(G(A_1), G(A_2), \dots, G(A_r))$ is distributed as a Dirichlet distribution with parameters $(\gamma H(A_1), \gamma H(A_2), \dots, \gamma H(A_r))$.

A measure drawn from a DP is almost surely discrete and has infinite dimensions. Sethuraman made this property explicit in its *stick-breaking representation* [13]. Specifically, if $G \sim \mathrm{DP}(\gamma, H)$, then $G = \sum_{k=}^{\infty} \beta_k \delta_{\phi_k}$ where $\phi_k \overset{\text{iid}}{\sim} H$ and $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$. $\boldsymbol{\beta}$ is the infinite dimensional vector of atom weights obtained from stick-breaking process $\beta_k = v_k \prod_{l=1}^{k-1}(a - v_l)$ and $v_l \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \gamma)$.

Instead of referring to G, a DP can often be viewed by samples drawn from G. This view is known as *Pólya urn scheme* [3]. If $G \sim \mathrm{DP}(\gamma, H)$ and $\theta_1, \dots, \theta_i \sim G$, after marginalizing out the random measure G the conditional predictive distributions have the following form:

$$\theta_i | \theta_{1:i-1}, \gamma, H \sim \sum_{i=1}^{i-1} \frac{1}{i-1+\gamma} \delta_{\theta_i} + \frac{\gamma}{i-1+\gamma} H \qquad (1)$$

As the results of the positive reinforcement effect of Equation 1 and G being discrete, $\{\theta_1, \dots, \theta_{i-1}\}$ take K distinct values where $K \leq i-1$. This shows the clustering property of DP. If we define $\{\phi_1, \dots, \phi_K\}$ to be the K unique atoms drawn from H and n_k to be the number of $\theta_{i'}$ equal to ϕ_k for $1 \leq i' < i$, Equation 1 can be re-expressed as Equation 2. This view is known as *Chinese restaurant process* (CRP).

$$\theta_i | \theta_{1:i-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{n_k}{i-1+\gamma} \delta_{\phi_k} + \frac{\gamma}{i-1+\gamma} H \qquad (2)$$

B. Dirichlet process mixture model

If we assume $G|\gamma, H \sim \mathrm{DP}\left(\gamma, H\right)$, then G is a discrete measure with infinite atoms; This is a desirable prior for density estimation problems with mixture modelling. By endowing the parameter space of mixture components with a DP prior, we induce infinitely many mixture components and therefore, infinitely many clusters. When coupled with appropriate data generation likelihood distribution F, the model is referred to as *Dirichlet process mixture model* (DPM) and formally defined as follows:

$$\begin{array}{rcl} G|\gamma, H & \sim & \mathrm{DP}\left(\gamma, H\right) \\ \theta_i|G & \overset{\text{iid}}{\sim} & G \\ x_i|\theta_i & \overset{\text{iid}}{\sim} & F\left(\theta_i\right) \end{array}$$

Using the equivalent CRP metaphor, we assume there exists a restaurant with an infinite number of tables. Customer θ_i enters the restaurant and selects table k with a probability proportional to the number of customers sitting at that table i.e. n_k and shares their dish ϕ_k, or selects a new table with a probability proportional to γ and orders a new dish from H as in Equation 2. The table assignments impose a partitioning on the customers. This result into clustering the data $\{x_i\}_{i=1}^{n}$ where each table is a cluster and the data for cluster k is generated from the likelihood distribution F with parameter ϕ_k.

C. Temporal Dirichlet process mixture model

As described in Section III-B, DPM is suitable for non-parametric clustering of exchangeable data. However, in many real-world problems, the data is not fully exchangeable as it arrives in multiple epochs. In this case, a useful approximation is assuming partial exchangeability, where the data arrives in T consecutive epochs. The epochs are not exchangeable, but the data within each epoch is exchangeable. It is desired to model the data such that the clusters are shared across epochs. Since a basic DP prior can not accommodate multiple epochs, DPM is not capable of modelling such behaviour. Moreover, some clusters may disappear over time, or new clusters emerge.

Ahmed and Xing [2] addressed this limitation by fitting a DPM on each epoch. Furthermore, they assumed the prior distribution on the parameter space of mixing components at time t, has the same structure of the posterior distribution at time $t-1$. Due to the conjugacy of the DP, the posteriors are also DP and therefore this model allows atoms drawn from the base measure to be shared across multiple epochs where atoms can disappear in any epoch, or new atoms can emerge. This model is called temporal DPM and formally defined as:

$$\begin{array}{rcll} G_1 & \sim & \mathrm{DP}\left(\gamma, H\right) & \\ G_t|G_{t-1}, \boldsymbol{\theta}_{t-1}, \gamma, H & \sim & \mathrm{DP}\left(\gamma + n_{t-1}, H_t\right) & 2 \leq t <\leq T \\ \theta_{t,i}|G_t & \overset{\text{iid}}{\sim} & G_t & \\ x_{t,i}|\theta_{ti} & \overset{\text{iid}}{\sim} & F\left(\theta_{t,i}\right) & \end{array}$$

where

$$H_t = \sum_{k \in I_{t-1}} \frac{n_{t-1,k}}{n_{t-1}+\gamma}\delta_{\phi_k} + \frac{\gamma}{n_{t-1}+\gamma}H$$

and H_t is the posterior of mixing component parameters at time t having observed $\boldsymbol{\theta}_{t-1} = \{\theta_{t-1,i}\}_{i=1}^{n_{t-1}}$ and $\theta_{t-1,i} \in \{\phi_k\}_{k=1}^{K}$.

Temporal DPM can be described as a generalization of CRP using a somewhat similar metaphor. This view is called *recurrent Chinese restaurant process* (RCRP) and depicted in Figure 1. In this view, there exists a restaurant with an infinite number of tables which operates in T days. Customers who enter on any given day, leave the restaurant at the end of the day. At any given day, a customer i i.e. $\theta_{t,i}$ enters the restaurant, selects a table, and enjoys a dish. If the table is occupied by other customers, $\theta_{t,i}$ shares their dish, otherwise, a new dish is ordered from the menu H.

While selecting a table for customer $\theta_{t,i}$, three scenarios might happen as shown in Equation 3 and Figure 1:

- $\theta_{t,i}$ selects a retained table from the previous day and enjoys the same dish that customers of that table ordered the day before (part ⓐ in Equation 3 and $\theta_{2,1}$ in Figure 1),
- $\theta_{t,i}$ joins a table which was instantiated in day t and share the dish (part ⓑ in Equation 3 and $\theta_{2,5}$ in Figure 1),
- or $\theta_{t,i}$ starts a new table and orders a new dish form the menu (part ⓒ in Equation 3 and $\theta_{2,3}$ in Figure 1).

RCRP is defined by:

$$\theta_{t,i}|\theta_{t-1,1:n_{t-1}}, \theta_{t,1:i-1}, H, \gamma \sim \tfrac{1}{n_{t-1}+i-1+\gamma} \times \left[\underbrace{\sum_{k \in I_{t-1}} \left(n_{t-1,k} + n_{t,k}^{(i)}\right)\delta_{\phi_k}}_{\text{ⓐ}} + \underbrace{\sum_{k \in I_t^{(i)} - I_{t-1}} n_{t,k}^{(i)}\delta_{\phi_k}}_{\text{ⓑ}} + \underbrace{\gamma H}_{\text{ⓒ}} \right] \quad (3)$$

where n_t denotes the number of customers at time t, $n_{t,k}$ denotes the number of customers at time t sitting at table k and enjoying dish ϕ_k, and $n_{t,k}^{(i)}$ represents the number of customers at time t sitting at table k and enjoying dish ϕ_k up to arrival of i^{th} customer. The set of tables occupied by customers at time t is shown by I_t.

D. Hierarchical Dirichlet process mixture model

The DPM is suitable for non-parametric clustering of exchangeable data in a *single* group of observations. However, many real-world problems are more appropriately modelled as comprising multiple groups of exchangeable data. In such cases, we can associate each group with a mixture model. But it is desirable to model the observations of different groups jointly to allow them to share their statistical strength. In Bayesian setting, this behaviour is naturally achieved using a hierarchical structure for prior.

To address this need, Hierarchical Dirichlet Process (HDP) offers an attractive extension to the basic DP by assuming the base measure of the DP is itself drawn from another DP, hence the name [15]. The key idea comes from the discreteness

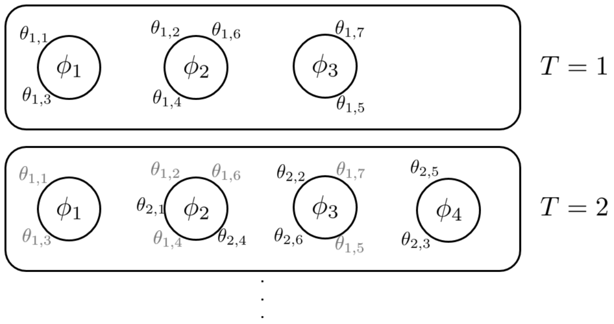


Figure 1: A depiction of recurrent Chinese restaurant process for two epochs. Customers ($\theta_{t,i}$s) are seated at tables (circles) in the restaurant and at each table a dish is served. At time t, customer i i.e. $\theta_{t,i}$ selects where to sit with a probability proportional to the collective number of customers at tables on days $t-1$ and t (Equation 3).

of DP which allows explicit statistical sharing across groups. Endowing the parameter space of a mixture model with an HDP prior results into a model that is ideal for modelling groups of data which are exchangeable within and across groups. It is called Hierarchical Dirichlet Process mixture model, or for short HDPMM, and formally defined as follows:

$$\begin{aligned} G_0|\gamma, H &\sim \mathrm{DP}(\gamma, H) \\ G_j|\alpha, G_0 &\overset{\text{iid}}{\sim} \mathrm{DP}(\alpha, G_0) \\ \theta_{ji}|G_j &\overset{\text{iid}}{\sim} G_j \\ x_{ji}|\theta_{ji} &\overset{\text{iid}}{\sim} F(\theta_{ji}) \end{aligned} \tag{4}$$

While each group j is modelled with a DPM, the group level base measure G_0 is drawn from a corpus level DP to allow different groups share their mixing components. G_0 being almost surely a discrete probability distribution, the children random measures G_j(s) necessarily share the same atoms across j and the mixture models in the different groups share mixture component as desired. However, the random distributions G_j are different.

Similar to CRP metaphor, there is an equivalent view of HDP which is the result of integrating out random measures. This view is called the *Chinese Restaurant Franchise* (CRF) and extends the CRP to allow multiple restaurants which share a set of dishes from a global menu [15]. Customer i in restaurant j i.e. θ_{ji} selects a table and enjoys a dish. θ_{ji} selects table b following a CRP as in Equation 5. If b is occupied by other customers, θ_{ji} shares their dish ψ_{jb} otherwise, a new dish is drawn from the global menu following another CRP as in Equation 6.

$$\theta_{ji}|\theta_{j1:i-1}, \alpha, G_0 \sim \sum_{b=1}^{B_j} n_{jb}\delta_{\psi_{jb}} + \alpha G_0 \tag{5}$$

$$\psi_{jb}|\psi_{11}, \psi_{12}, \ldots, \psi_{jb-1}, \gamma, H \sim \sum_{k=1}^{K} m_k \delta_{\phi_k} + \gamma H \tag{6}$$

where n_{jb} is the number of customers sitting at table b in restaurant j, B_j is the number of active tables in restaurant j, ψ_{jb} is the dish being served at table b in restaurant j, and m_k is the total number of tables serving dish k in all restaurants.

E. Infinite dynamic topic model

In many problems, the data arrives in multiple epochs where each epoch consists of multiple groups of observations. It is desired to model epochs as comprising multiple groups that share their clusters over the groups and the epochs. Section III-C reviewed a model for partially exchangeable data where the data is represented in epochs, exchangeable within epochs but not across them. It assumes each epoch is flat and can not model hierarchy of observations in epochs. On the other hand, Section III-D reviewed an approach for hierarchical data, modelling multiple groups within a single epoch. But it can not accommodate multiple epochs. Combining the two results in a suitable model for non-parametric clustering of partially exchangeable groups of observations where the data is divided into T consecutive epochs. The epochs are not exchangeable, but the data within each epoch is assumed to be exchangeable within and across groups.

Ahmed and Xing [**?**] addressed this problem by fitting an HDPMM on each epoch and assuming the prior distribution on the parameter space at time t has the same structure of the posterior distribution of the parameters at time $t-1$. They achieved this by substituting the top level DP of an HDP with a temporal DP. This way the atoms drawn from the top level base measure:

1) are shared across epochs,
2) are shared across groups within each epoch,
3) can emerge or disappear over time.

This model is called *infinite dynamic topic model* (iDTM) ,and formally describe iDTM as the following. We shorten the notations with three indices by dropping the comma, for instance, $\theta_{t,j,i}$ and θ_{tji} refer to the same thing.

$$\begin{aligned} G_0^1|\gamma, H &\sim \mathrm{DP}(\gamma, H) \\ G_0^t|\boldsymbol{\theta}_{t-1}, \gamma, H &\sim \mathrm{DP}(\gamma + m_{t-1}, H_t) \quad 2 \le t \le T \\ G_j^t|\alpha, G_0^t &\sim \mathrm{DP}\left(\alpha, G_0^t\right) \\ \theta_{tji}|G_j^t &\overset{\text{iid}}{\sim} G_j^t \\ x_{tji}|\theta_{tji} &\overset{\text{iid}}{\sim} F(\theta_{tji}) \end{aligned} \tag{7}$$

where

$$H_t = \sum_{k \in I_{t-1}} \frac{m_{t-1,k}}{m_{t-1} + \gamma}\delta_{\phi_k} + \frac{\gamma}{m_{t-1} + \gamma} H$$

is the posterior of mixing components parameters at time t having observed $\boldsymbol{\theta}_t = \left\{\{\theta_{tji}\}_{i=1}^{N_{tj}}\right\}_{j=1}^{N_t}$ and $\theta_{tji} \in \{\phi_k\}_{k=1}^{K}$, $m_{t,k}$ is the number of tables that serve dish k in all restaurants at day t, and $m_t = \sum_{k \in I_t} m_{t,k}$ as become clearer later.

This model becomes clearer using a metaphor similar to CRP, depicted in Figure 2. It is called *recurrent Chinese restaurant franchise* (RCRF) and achieved by integrating out the random measures in Equation 7. Each epoch is modelled

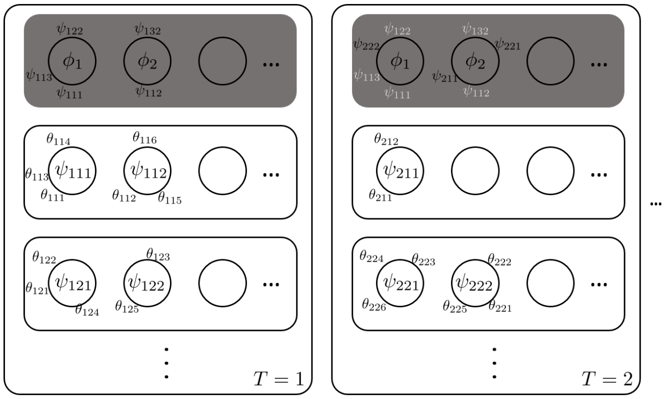


Figure 2: A depiction of recurrent Chinese restaurant franchise for two epochs. Each epoch is modelled with a CRF. Shadeless inner rectangles represent restaurants. Customers (θ_{tji}) are seated at tables (circles) in the restaurants and at each table a dish is served. The dish is served from a global menu (ϕ_k) where the parameter ψ_{tjb} is a table-specific indicator that serves to index items on the global menu. Illustrated as shaded rectangles, the parameters ψ_{tjb} are modelled with a recurrent Chinese restaurant process. In other words, tables at day t, serve dish k with a probability proportional to the collective number of tables that serve dish k in days t and $t-1$, hence the washed out table-specific indicators in the shaded rectangle. The customer θ_{tji} sits at the table to which it has been assigned in Equation 8 and enjoys the dish to which he was assigned in Equation 9.

with a CRF as explained in Section III-D and then the epoch-specific global menus are modelled with RCRP as explain in Section III-C. The first epoch is modelled exactly as in the CRF. At epoch t, customer i in restaurant j i.e. θ_{tji} selects a table and enjoys a dish. θ_{tji} selects table b with a probability proportional to the number of customers occupying it and selects a new table with a probability proportional to α as in Equation 8. If the table is occupied, they share the dish ψ_{tjb} otherwise, a new dish is sampled from the epoch-specific global menu as in Equation 9.

$$\theta_{tji}|. \sim \sum_{b=1}^{B_{tj}} n_{tjb}\delta_{\psi_{tjb}} + \alpha\delta G_0^t \tag{8}$$

$$\psi_{tjb}|. \sim \sum_{k\in I_{t-1}} \left(m_{t-1,k} + m_{tk}\right)\delta_{\phi_k} + \sum_{k\in I_t - I_{t-1}} m_{tk}\delta_{\phi_k} + \gamma H \tag{9}$$

where n_{tjb} denotes the number of customers at table b in restaurant j at day t, B_{tj} is the number of active tables in restaurant j at day t, ψ_{tjb} represents the dish being served at table b in restaurant j at day t, and m_{tk} is the total number of tables in day t in all restaurants that serve dish ϕ_k.

As the formal definition and the RCRF view show, after obtaining the epoch-specific global menu (G_0^t), the rest is identical to HDPMM. Therefore, groups in each epoch, share their atoms. The atoms are also shared across epochs as the epoch-specific global menus are modelled with a temporal DP.

TDPM and iDTM as introduced in [2], [**?**] are evolutionary topic models. They model the clusters in the data as well as their changes over time. This is a desirable behaviour in social media and news data. But in the scientific literature, changes in topics usually result in new topics to emerge. Moreover, considering the topic evolution severely affects the identifiability of the model. For example, if we assume a topic at time t such as ϕ_t shifts to a topic in time $t+1$ as $\phi_{t+1} = \phi_t + \epsilon$, then the model will have difficulty in identifying ϕ_{t+1} as an entirely new topic where ϕ_t is dead, or as the continuum of ϕ_t with evolution. To address these issues, we fix the topics in our framework. This way all epochs share the same set of topics and changes in topics are modelled as new topics as desired.

1) Inference : Using the RCRF view, the sampling is straightforward. The state space involves table associations for customers $\mathbf{b}$, dish associations for tables $\mathbf{k}$, and the dishes $\boldsymbol{\phi}$. Here, b_{tji} represents the table associated with customer i, in restaurant j at epoch t; k_{tjb} represents the dish associated with table b, in restaurant j at epoch t; and ϕ_k represents dish k. Adding a superscript to a variable represents the same quantity without the contribution of the superscript.

a) Sampling tables: To compute the conditional distribution of table associations for each customer, we use Equation 8

$$p\left(b_{tji} = b|\mathbf{b}_{tj}^{-i}, \mathbf{k}_{t-1,t}, \boldsymbol{\phi}, x_{tji}\right) \propto$$

$$\begin{cases} n_{tjb}^{-tji} f_{k_{tjb}}^{-x_{tji}}\left(x_{tji}\right) & b \text{ is used} \\ \alpha p\left(x_{tji}|\mathbf{b}_{tj}^{-i}, b_{tji} = b^{new}, \mathbf{k}\right) & b \text{ is new} \end{cases}$$

If the sampled value of b_{tji} is b^{new}, then we obtain a sample of $k_{tjb^{new}}$ by sampling from Equation 9

$$p\left(k_{tjb^{new}} = k|\mathbf{k}_{t-1,t}, \mathbf{b}\right) \propto$$

$$\begin{cases} \left(m_{t-1,k} + m_{t,k}^{-tjb}\right) f_k^{-x_{tji}}\left(x_{tji}\right) & k \text{ is used} \\ \gamma f_{k^{new}}^{-x_{tji}}\left(x_{tji}\right) & k \text{ is new} \end{cases}$$

b) Sampling dishes: Since changing the dish which is served at a table affects all customers sitting at that table, we should consider all of the customers in the likelihood. In other words:

$$p\left(k_{tjb} = k|\mathbf{k}_{t-1,t}, \mathbf{b}\right) \propto$$
$$\begin{cases} \left(m_{t-1,k} + m_{t,k}^{-tjb}\right) f_k^{-\mathbf{x}_{tjb}}\left(\mathbf{x}_{tjb}\right) & k \text{ is used} \\ \gamma f_{k^{new}}^{-\mathbf{x}_{tjb}}\left(\mathbf{x}_{tjb}\right) & k \text{ is new} \end{cases}$$

F. Popularity of topics

We use a post hoc calculation, based on the observed probability of each topic given the epoch to find the popularity of the topic over time. As the result, we relax the constraints that models such as [18], [5] impose on the model. We assume z_{tji} indicates the mixture component to which x_{tji} belongs and define $\hat{p}\left(k|t\right)$ as the empirical probability of topic k in an arbitrary group in time t:

$$\begin{aligned}\hat{p}\left(k|t\right) &= \sum_{j=1}^{N_t} \hat{p}\left(k|\mathbf{x}_{tj}\right)\hat{p}\left(\mathbf{x}_{tj}|t\right) \\ &= \frac{1}{N_t}\sum_{j=1}^{N_t} \hat{p}\left(k|\mathbf{x}_{tj}\right) \qquad (10) \\ &= \frac{1}{N_t}\sum_{j=1}^{N_t}\sum_{i=1}^{N_{tj}} I\left(z_{tji}=k\right)\end{aligned}$$

IV. Experiments and results

In this section, we report our results of conducting experiments on a corpus of scholarly literature related to ASD. ASD is a life-long complex neurodevelopmental disorder with poorly understood aetiology on the one hand, and without treatment or medication to cure its core symptoms on the other hand. Affecting approximately 1.5% of the population, it is characterised by severe impairments in social interaction, communication, and in some cases cognitive abilities. Considering the prevalence, symptoms, and its social and economic burden, it is not surprising that it has been attracting an increasing amount of research attention. This has resulted in a rapid growth of a broad corpus of literature with sometimes hypotheses with little evidence and even in direct contradiction.

First, we describe our data collection, then show the usefulness of our approach using two case studies, and finally, describe our additional contribution that comes in the form of a free online tool that we developed to aid ASD researchers.

A. Data collection and pre-processing

We collected a comprehensive dataset related to ASD using the API provided by Scopus[3]. We assumed a paper is related to ASD if the term "autism" is present in its title, abstract, or indexed keywords by its author(s). Even though the earliest publication that fits these criteria was published in 1946 [8], the earliest relevant paper indexed by Scopus goes back to 1977. Comparing the results of querying Scopus with PubMed shows there exists 230 article published between 1946 - 1977 that are not indexed by Scopus. These papers are scattered in a duration of 30 years and disregarding them does not make a significant difference in the final results.

Querying Scopus yields a corpus of 46,218 publications spanning from 1977 to 2015. The distribution of publications in the corpus over time is visualised in Figure 3. As the figure shows, there has been a three-fold increase in the number of published papers per annum in the past decade.

To evaluate our method, we use the title and abstract of publications. For each article, we create a new document by concatenating its title and abstract and fit the model to the corpus of the newly generated documents. In order to reduce the dimensionality and the effect of inflexions of words, we perform a soft lemmatization based on WordNet lexicon on the corpus, followed by removing the so-called stop-words. Then we construct a vocabulary by selecting a subset of the most frequent terms such that it explains 90% of the energy of the corpus. This results in a vocabulary of 4,763 terms and a corpus of 3,735,764 tokens in total. Figure 4 shows the length distribution of documents in our corpus.

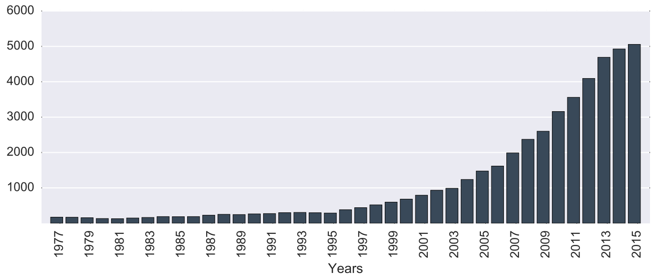


Figure 3: The rapid rise in the rate of publications concerned with ASD. Shown is the number of publication indexed by Scopus and matching the criteria described in Section IV-A per year. There has been a three-fold increase in the number of publications per annum in the past decade.

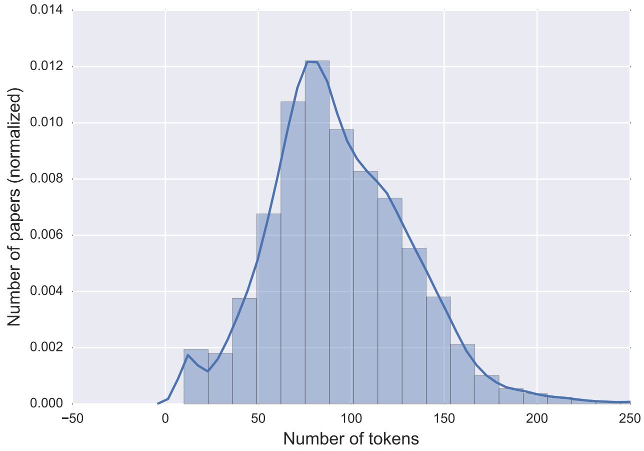


Figure 4: The length distribution of documents of the ASD corpus after using the vocabulary to represent the data. On average, each document is constructed of 94 tokens.

B. Experiment

We assume each epoch is one year and use the model explained in Section III-E to obtain the topics of the corpus described in the previous section. We collect 500 samples for burn-in, allowing the Gibbs chain to converge and 2000 samples after the burn-in period. In total, 53 topics are found. In the next, section we will analyse two of them in detail.

In each case study, we carefully examine the topics associated with a major underlying theme. We show our framework is capable of discovering various aspects that exist to an overarching theme. For example, we not only find an underlying theme associated with ASD and genetics, but we also find more fine-grained strands of this theme in the literature, such as gene mapping, genetic mutation, and other topics related to genetics.

[3] https://www.scopus.com/

C. Case study 1: ASD and vaccination

In 1998, Wakefield et al. published a work which reported epidemiological findings linking Measles, Mumps, and Rubella (MMR) vaccination and the development of ASD and colitis [16]. Despite the full retraction of the article, after numerous subsequent studies failed to show the claimed link, vaccination continues to be an occurring theme in ASD literature.

We begin by identifying the topic(s) with the highest probability of the terms "vaccine" and "vaccination" conditioned on the topic. Two topics are found. The first, depicted in Figure 5a as its most probable words suggest, is associated with MMR vaccination. In other words, it reflects the effect of the falsely claimed causal link between MMR vaccination and the development of ASD on the literature, as Wakefield's paper gave rise to numerous studies investigating the claimed link. Looking at some of the most contributing papers to this topic agrees with this explanation:

- Afzal, M. A., P. A. Pipkin, and P. D. Minor. "Absence of chicken myelin basic protein residues in commercial formulations of MMR vaccine." Vaccine 19.4 (2000): 442-446.
- Edwards, Carolyn. "Is the MMR vaccine safe?." Western Journal of Medicine 174.3 (2001): 197.
- Singh, Vijendra K., et al. "Abnormal measles-mumps-rubella antibodies and CNS autoimmunity in children with autism." Journal of biomedical science 9.4 (2002): 359-364.
- Mehta, Bijal K., and Kerim M. Munir. "Does the MMR vaccine and secretin or its receptor share an antigenic epitope?." Medical hypotheses 60.5 (2003): 650-653.

However, this topic did not survive until now. It appeared in 2000 and disappeared in 2009. Instead, another topic associated with vaccination was born in 2005 that continues to exist until this day. But unlike the previous topic, it is not specifically about MMR. Looking at its word-cloud (Figure 5b) and most related papers in different years reveal that it is more related to the safety monitoring of vaccination and immunization. Some of the papers that mostly contributed to this topic are:

- Lee, C. J., et al. "Safety monitoring in vaccine development and immunization." ACTA PAEDIATRICA TAIWANICA 47.1 (2006): 7.
- Carbarns, Ian RI, et al. "Use of risk management planning to enhance safety of medicines." International Journal of Pharmaceutical Medicine 21.6 (2007): 415-426.
- Iskander, John K., et al. "Selected major issues in vaccine safety." Annales Nestlé (English ed.) 66.2 (2008): 93-102.
- Hammer, Lawrence D., et al. "Increasing immunization coverage." Pediatrics 125.6 (2010): 1295-1304.

1) Vaccination and exposure to environmental factors: Investigating the results reveals another topic that at first might seem unrelated to vaccination, but some understanding of the domain shows how the two are connected.

As mentioned earlier, Wakefield suggested that MMR vaccine contributes to the developing ASD [16] and many subsequent works proved him to be wrong. But, this controversial topic gave rise to many studies that investigated the contribution of environmental factors to the developing ASD; from prenatal exposure to valproic acids to postnatal exposure to mercury. This is another concept that our framework has captured.

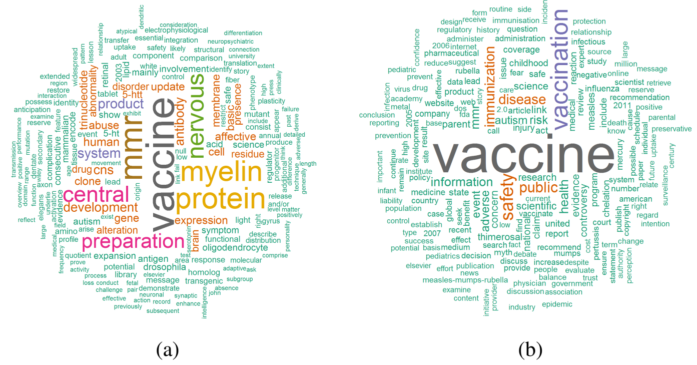


Figure 5: An illustration of topics discovered by the proposed method in the ASD corpus and related to vaccination. (a) is related to the falsely claimed link between MMR vaccination and developing ASD, and (b) is related to safety monitoring of vaccination and medicine

This topic is associated with the exposure of the developing brain to the environmental factors and their effects on its development. It appears in 2000 and lives until today. This is consistent with our findings about vaccination topic explained in Section IV-C as it also appeared in 2000. Figure 6a depicts the word-cloud representation of this topic.

D. Case study 2: ASD and genetics

Although the exact aetiology of ASD is still poorly understood, genetic factors are clearly implicated in the causation of ASD. Children that have siblings with ASD have higher risks of pervasive developmental disorder. Several twin studies suggest that this aggregation within families is best explained by shared genes as opposed to shared environment. In twins that one child has ASD, if they are identical the other child will be affected 95% of the time and 31% of the time if they are non-identical [14]. These results have led to a tremendous effort in research to try to unravel the genetic factors underlying the disorder. Therefore, genetic factors as a major theme of research in ASD is a good case study on which we can illustrate the usefulness of our proposed method.

Similar to Section IV-C, we start by identifying the topics with the highest probability of the terms "gene" or "genetic" conditioned on the topic. Three topics are found.

The first as depicted in Figure 6b, is associated with gene mapping and identifying the loci of the genes that increase the risk of ASD. It appears in 1985 and survives until now. Some of the articles that mostly contributed to this topic are:

- Spence, M. Anne, et al. "Gene mapping studies with the syndrome of autism." Behavior genetics 15.1 (1985): 1-13.

- Gurling, Hugh. "Candidate genes and favoured loci: strategies for molecular genetic research into schizophrenia, manic depression, autism, alcoholism and Alzheimer's disease." Psychiatric developments 4.4 (1985): 289-309.
- Chodirker, B. N., et al. "Fragile 19p13 in a family with mental illness." Clinical genetics 31.1 (1987): 1-6.

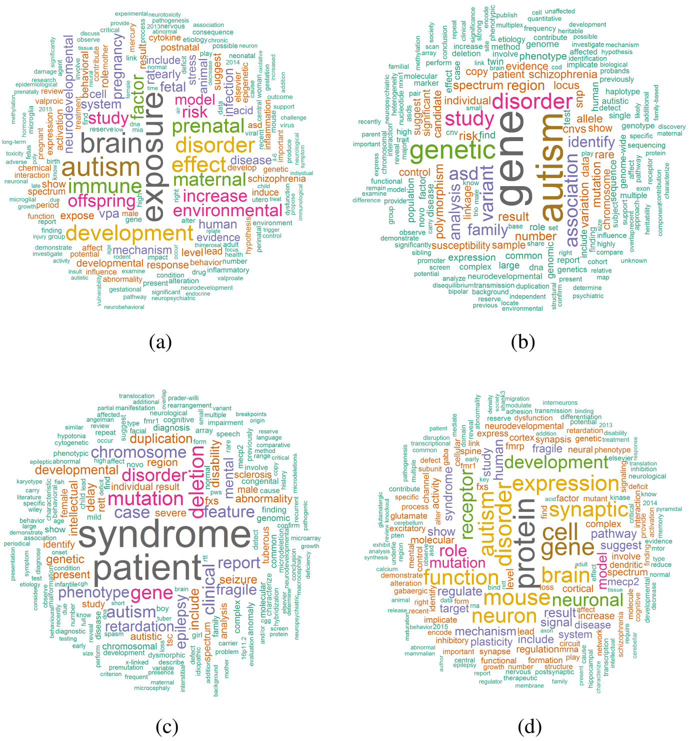
The second topic starts from 1986 and as Figure 6c shows, it is mainly about gene mutation and other neurodevelopmental disorders similar to ASD that are caused by gene mutation, for example, the Rett syndrome and Fragile X. Rett Syndrome is a rare genetic disorder with symptoms that can be most easily confused with those of ASD. It is caused by the mutation in the MeCP2 gene. Fragile X syndrome is caused by the mutation of the FMR1 gene and nearly half of all children with fragile x meet the criteria of diagnosis of autism. A few of most relevant papers include:

- Goutières, Françoise, et al. "Atypical forms of Rett syndrome." American Journal of Medical Genetics 25.S1 (1986): 183-194.
- Edwards, D. R., et al. "Autism in association with fragile X syndrome in females: implications for diagnosis and treatment in children." Neurotoxicology 9.3 (1987): 359-365.
- Fleury, P., M. Van Schooneveld, and J. W. Delleman. "Neurological, ophthalmological and nephrological aspects of tuberous sclerosis." Tuberous Sclerosis and Neurofibromatosis: Epidemiology, Pathophysiology, Biology and Management. Amsterdam: Elsevier (1990): 211-226.

The third topic is about investigations of gene expression and synaptic plasticity through a cell biology lens. Gene expression is tightly linked to our understanding of proteins and synaptic plasticity is highly affected by the alteration of the number of neurotransmitter receptors which are proteins, hence the big protein in the centre of Figure 6d. Moreover, many of the risk genes that have been linked to the ASD (such as MeCP2) encode synaptic scaffolding proteins and receptors. Synaptic plasticity is an important neurochemical foundation of learning and memory, impairments of which are symptoms of ASD. Many studies have identified several risk genes that are key regulators of synaptic plasticity.

This topic appeared in 2000 and survives until now. Some of the most contributing papers to this topic agrees with our explanation:

- Court, Jenny A., et al. "Nicotinic receptors in human brain: topography and pathology." Journal of chemical neuroanatomy 20.3 (2000): 281-298.
- Akbarian, Schahram, et al. "Expression pattern of the Rett syndrome gene MeCP2 in primate pre-frontal cortex." Neurobiology of disease 8.5 (2001): 784-791.
- Lee, M., et al. "Nicotinic receptor abnormalities in the cerebellar cortex in autism." Brain 125.7 (2002): 1483-1495.



(a)

(b)

(c)

(d)

Figure 6: An illustration of topics discovered by the proposed method in the ASD corpus and related to contributing factors to the causation of ASD. (a) is related to the environmental contribution, (b) is related to gene mapping and discovering the loci of genes that increase the risk of ASD, (c) is related to gene mutation and disorders similar to ASD that it can cause, and (d) is about risk genes that are linked to alteration of synaptic plasticity and therefore contribute to the developing of ASD.

E. Analysis of research interest

Given the space of topics defined in Section IV-B, we now examine the history and popularity of them in the entire corpus from 1977 to 2015 using Equation 10. We select to show the popularity of topics associated with genetics and ASD that previously were explained in Section IV-D. Figure 7 shows the changes in research interest over time to these topics.

As the Figure 7 shows, the topic about gene expression is relatively newer than others and has seen a sharp increase in interest from the time it appeared. However, The other two are not *hot* any more. The interest in gene mutation has consistently declined over the past 15 years and the popularity of gene mapping fluctuates over the past 10 years, making it a steady topic.

Using this method, we can find the topics that get an increasing amount of attention form the research community. For example, the following two topics are fairly new and on the rise:

- the use of new technologies such as virtual reality or robots in therapy, and
- the use of computational methods in studying EEG and fMRI signals of children with ASD

Similarly, we can find the topics in which researches have lost their interest over time; for instance, the theory of mind. It refers to the notion that many individuals with ASD,

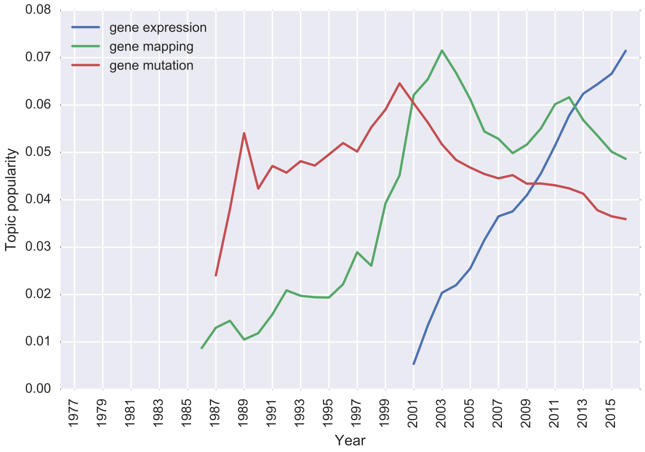


Figure 7: Research interest in topics associated with genetics in the ASD corpus over time. For an explanation of these topics see Section IV-D. The topic about gene expression receiving an increasing amount of research interest over time. In other words, gene expression is a *hot* topic in ASD domain. On the other hand, the popularity of the other two has decreased over the past decade.

have difficulties understanding other people's attitudes and emotions.

F. Topic browser

To provide access to our results to the research community interested in ASD, we prepared an interactive visualisation[4]. It facilitates analysing the topics that our framework discovered in the corpus and allows researchers to search through 40 years of publications by topic. Figure 8 shows a screenshot of the visualisation. The inner ring represents the topics, the middle ring shows the top 10 most probable of each topic, and the outer ring shows the most relevant papers to a selected topic from 2005 - 2015.

V. Conclusion

We proposed a framework capable of discovering the underlying topical structure of the research domain, and tracking the lifetime and popularity of them over time. We assume the data consists of T epochs and each epoch comprises of multiple groups of observations. The data is exchangeable within each epoch but not across them and within each epoch, the data is exchangeable within each group but not across the groups. We used the recurrent Chinese restaurant franchise with static topics to model the topical structure of the data and their lifetime and using post-processing, we found the popularity of topics in the research community over time.

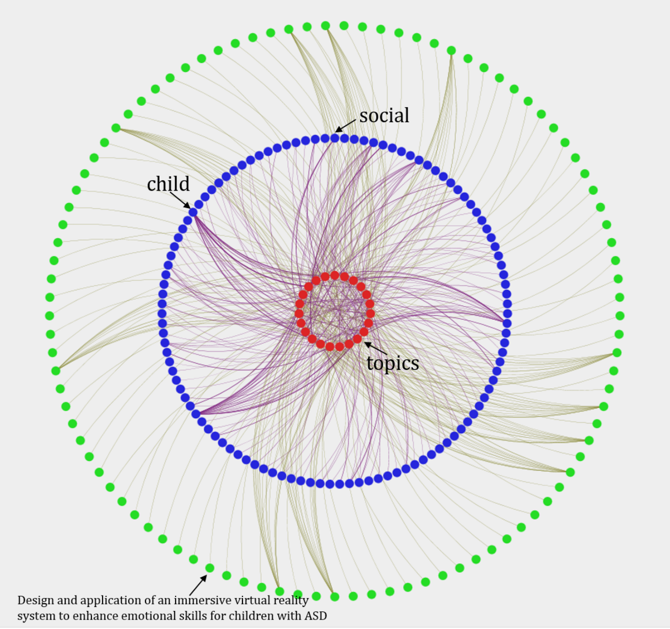

social
child
topics
Design and application of an immersive virtual reality
system to enhance emotional skills for children with ASD

Figure 8: A visualization of the topics discovered by the proposed method in the ASD corpus. The inner ring represents the topics, the middle rings shows the union of top 10 most frequent words of the topics, and the outer ring is the union of the most relevant paper to each topic from 2005 to 2015. Papers are represented by their unique Scopus ID. The online version is interactive and includes search.

We demonstrated the power of our proposed framework on a dataset of ASD-related medical literature with nearly 50,000 publications which spans over 40 years. We investigated two important research themes in ASD and showed that our framework is capable of discovering the fine-grained topics related to these themes, as well as their lifetime and their popularity. The results are made available to ASD researchers in the form of an online interactive visualisation. The source code of our model is also freely available for those who are interested in using it for other data collections.

References

[1] , "————," pp. 550–562.

[2] A. Ahmed and E. P. Xing, "Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering." in *SDM*. SIAM, 2008, pp. 219–230.

[3] D. Blackwell and J. B. MacQueen, "Ferguson distributions via pólya urn schemes," *The annals of statistics*, pp. 353–355, 1973.

[4] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 113–120.

[5] A. Dubey, A. Hefny, S. Williamson, and E. P. Xing, "A nonparametric mixture model for topic modeling over time." in *SDM*. SIAM, 2013, pp. 530–538.

[6] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[7] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 363–371.

[8] L. Kanner, "Irrelevant and metaphorical language in early infantile autism," *American journal of Psychiatry*, vol. 103, no. 2, pp. 242–246, 1946.

[9] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models:\# twitter trends detection topic model online." in *COLING*, 2012, pp. 1519–1534.

[10] D. Mimno, "Computational historiography: Data mining in a century of classics journals," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 5, no. 1, p. 3, 2012.

[4] http://google.com - link removed due to the blind review process

[11] U. C. Priva and J. L. Austerweil, “Analyzing the history of cognition using topic models,” *Cognition*, vol. 135, pp. 4–9, 2015.

[12] L. Ren, D. B. Dunson, and L. Carin, “The dynamic hierarchical dirichlet process,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 824–831.

[13] J. Sethuraman, “A constructive definition of dirichlet priors,” DTIC Document, Tech. Rep., 1991.

[14] H. Taniai, T. Nishiyama, T. Miyachi, M. Imaeda, and S. Sumi, “Genetic influences on the broad spectrum of autism: Study of proband-ascertained twins,” *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 147, no. 6, pp. 844–849, 2008.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes,” *Journal of the american statistical association*, vol. 101, no. 476, 2006.

[16] Wakefield, Andrew J and Murch, Simon H and Anthony, Andrew and Linnell, John and Casson, DM and Malik, Mohsin and Berelowitz, Mark and Dhillon, Amar P and Thomson, Michael A and Harvey, Peter and others, “Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children,” *The Lancet*, vol. 351, no. 9103, pp. 637–641, 1998.

[17] C. Wang, D. Blei, and D. Heckerman, “Continuous time dynamic topic models,” *arXiv preprint arXiv:1206.3298*, 2012.

[18] X. Wang and A. McCallum, “Topics over time: a non-markov continuous-time model of topical trends,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 424–433.

[19] J. Zhang, Y. Song, C. Zhang, and S. Liu, “Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1079–1088.

Appendix

In Section IV we investigated two case studies to show the power of our framework. We described the topics related to genetics and vaccination in the domain of ASD. In this appendix, we provide more details about those topics we discovered (and mentioned for instance in Section IV-E) but did not expand to keep the consistency of the text.

We discovered a topic related to data-driven methods in studying EEG and fMRI signals of those with ASD. It appeared in 2014 and its most probable words are data, method, analysis, model, network, asd, autism, feature, and study. The following papers contribute most to this topic:

- Suckling, John, et al. "Are power calculations useful? A multicentre neuroimaging study." Human brain mapping 35.8 (2014): 3569-3577.
- Matlis, Sean, et al. "Robust disruptions in electroencephalogram cortical oscillations and large-scale functional networks in autism." BMC neurology 15.1 (2015): 1.

We discovered a topic which is related to the use of technology in therapy and mostly for social interaction. It appeared in 2005 and survives until now. Most frequent words of this topic are child, autism, social, interaction, design, system, robot, study, paper, and technology. The following paper made the most contribution to this topic:

- Grynszpan, Ouriel, J-C. Martin, and Jacqueline Nadel. "Using facial expressions depicting emotions in a human-computer interface intended for people with autism." Intelligent Virtual Agents. Springer Berlin Heidelberg, 2005.
- Mohamed, A. Ould, et al. "Attention analysis in interactive software for children with autism." Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility. ACM, 2006.

We discovered a topic reflecting the affective involvement of the families and how they cope with the difficulties of having a family member on the spectrum. It appeared in 2007 and survives until now. The most frequent words of this topic consist of child, parent, autism, asd, sleep, disorder, mother, stress, problem, and study. The following papers made the most contribution to this topic:

- Montes, Guillermo, and Jill S. Halterman. "Psychological functioning and coping among mothers of children with autism: a population-based study." Pediatrics 119.5 (2007): e1040-e1046.
- Pottie, Colin G., and Kathleen M. Ingram. "Daily stress, coping, and well-being in parents of children with autism: a multilevel modeling approach." Journal of Family Psychology 22.6 (2008): 855.

Our framework discovered a topic associated with the deficits in the social behaviour and its relation to the level of oxytocin in mouse models of ASD. It appeared in 1996 and lives until the present day. The most frequent words of this topic are social, mouse, behaviour, model, oxytocin, autism, behavioural, deficit, animal, and human. The following list is a few of the most relevant papers to this topic:

- Winslow, James T., and Thomas R. Insel. "The social deficits of the oxytocin knockout mouse." Neuropeptides 36.2 (2002): 221-229.
- Crawley, Jacqueline N., et al. "Social approach behaviors in oxytocin knockout mice: comparison of two independent lines tested in different laboratory environments." Neuropeptides 41.3 (2007): 145-163.

We discovered a topic associated with the relation of food habits, nutrients, and diets with ASD. It appeared in 2007 and is still alive. The most frequent words of this topic are food, diet, ocd, vitamin, symptom, dietary, gastrointestinal, disease, obsessive-compulsive, and eat. The most contributing papers are as follows:

- Ho, Helena H., Linda C. Eaves, and Diana Peabody. "Nutrient intake and obesity in children with autism." Focus on Autism and Other Developmental Disabilities 12.3 (1997): 187-192.
- Cornish, E. "A balanced approach towards healthy eating in autism." Journal of Human Nutrition and Dietetics 11.6 (1998): 501-509.

We discovered a topic associated with anxiety and social dysfunction in adolescents with ASD. It appears in 2007 and lives until now. The most frequent words to this topic are asd, autism, disorder, social, spectrum, trait, anxiety, group, study, individual and a one of the most relevant papers is:

- Jamison, T. Rene, and Jessica Oeth Schuttler. "Examining social competence, self-perception, quality of life, and internalizing and externalizing symptoms in adolescent females with and without autism spectrum disorder: a quantitative design including between-groups and correlational analyses." Molecular autism 6.1 (2015): 1.