MY «ACT» HAS ENDED BY BECOMING AN INTEGRAL PART OF MY NATURE, I TOLD MYSELF. IT'S NO LONGER AN ACT.

YUKIO MISHIMA

ONLINE DISCUSSIONS THROUGH THE LENS OF INTERACTION PATTERNS

MATTIA SAMORY

SUPERVISORE
CH.MO PROF. ENOCH PESERICO

COORDINATORE DI INDIRIZZO
CH.MO PROF. CARLO FERRARI

DIRETTORE DELLA SCUOLA
CH.MO PROF. MATTEO BERTOCCO

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI INGEGNERIA
DELL'INFORMAZIONE

SCUOLA DI DOTTORATO DI RICERCA
IN INGEGNERIA DELL'INFORMAZIONE

INDIRIZZO SCIENZA E TECNOLOGIA
DELL'INFORMAZIONE - CICLO XXVIII

Abstract

Computer-mediated communication is arguably prevailing over face-to-face. However, many of the subtleties that make in-person communication personal, cues such as an ironic tone of voice or an effortless posture, are inherently impossible to render through a screen. The context vanishes from the conversation - what is left is therefore mostly text, enlivened by occasional multimedia. At least, this seems the dominant opinion of both industry and academia, that recently focused considerable resources on a deeper understanding of natural and visual language.

I argue instead that richer cues are missing from online interaction only because current applications do not acknowledge them – indeed, communication online is already infused with nonverbal codes, and the effort needed to leverage them is well worth the amount of information they carry. This dissertation therefore focuses on what is left out of the traditional definition of content: I refer to these aspects of communication as *content-agnostic*. Specifically, this dissertation makes three contributions.

First, I formalize what constitutes content-agnostic information in computer-mediated communication, and prove content-agnostic information is as personal to each user as its offline counterpart. For this reason, I choose as a venue of research the web forum, a supposedly text-based, impersonal communication environment, and show that it is possible to attribute a message to the corresponding author solely on the basis of its content-agnostic features – in other words, without looking at the content of the message *at all*.

Next, I display how abundant and how varied is the content-agnostic information that lies untapped in current applications.To this end, I analyze the content-agnostic aspects of one type of interaction, the quote, and draw conclusions on how these may support discussion, signal user status, mark relationships between users, and characterize the discussion forum as a community. One interesting implication is that discussion platforms may not need to introduce new features for supporting social signals, and conversely social networks may better integrate discussion by enhancing its content-agnostic qualities.

Finally, I demonstrate how content-agnostic information reveals user behavior. I focus specifically on trolls, malicious users that disrupt communities through deceptive or manipulative actions.

In fact, the language of trolls blends in with that of civil users in heated discussions, which makes collecting irrefutable evidence of trolling difficult even for human moderators. Nonetheless, I show that a combination of content-agnostic and linguistic features sets apart discussions that will eventually be trolled, and reactions to trolling posts. This provides evidence of how content-agnostic information can offer a point of view on user behavior that is at the same time different from, and complementary to, that offered by the actual content of the contribution.

Popular up and coming platforms, such as Snapchat, Tumblr, or Yik Yak, are increasingly abandoning persistent, threaded, text-based discussion, in favor of ephemeral, loosely structured, mixed-media content. Although the results of this dissertation are mostly drawn from discussion forums, its research frame and methods should apply directly to these other venues, and to a broad range of communication paradigms. Also, this is but a preliminary step towards a fuller understanding of what additional cues can or should complement content to overcome the limitations of computer-mediated communication.

Sommario

Interagiamo sempre più attraverso uno schermo, al costo di perdere tutti quei dettagli che caratterizzano la comunicazione di persona: un tono di voce ironico o una posa *nonchalant* sono impossibili da digitalizzare. Le conversazioni digitali si spogliano del contesto: quel che rimane è prevalentemente testo, arricchito al più dall'occasionale contenuto multimediale. Almeno, questa sembra essere l'opinione prevalente di industria ed accademia, le quali hanno concentrato le proprie attenzioni sull'estrarre significato da linguaggio scritto e visivo.

La mia tesi, invece, è che questi dettagli non siano presenti nelle nostre interazioni attraverso lo schermo solo perché non messi a frutto, e quindi nascosti, dalle attuali applicazioni – la comunicazione online è caratterizzata da un proprio linguaggio nonverbale, e la quantità di informazione che esprime ben ripagherebbe lo sforzo necessario per estrarla. Questa tesi si concentra su ciò che viene escluso dalla tradizionale definizione di contenuto: mi riferirò a questi aspetti della comunicazione come "agnostici rispetto al contenuto". Nel dettaglio, questa tesi porta tre principali contributi alla letteratura esistente.

Il primo è una formalizzazione di "agnostico rispetto al contenuto" nel contesto delle comunicazioni informatiche, ed una prova del fatto che le informazioni "agnostiche rispetto al contenuto" siano caratteristiche individuali, così come accade nel mondo fisico. Per far ciò, fornisco un'analisi delle comunicazioni su web forum, una piattaforma di comunicazione considerata prevalentemente impersonale e testuale, e dimostro che è possibile identificare l'autore di un messaggio usando esclusivamente informazioni "agnostiche rispetto al contenuto" – in altre parole, senza leggere il messaggio.

Il secondo contributo è una dimostrazione del fatto che le attuali applicazioni per comunicare tramite schermo ignorino una quantità e varietà di informazioni "agnostiche rispetto al contenuto", e che queste abbiano significato convenzionale. A tal fine concentro i miei studi su una particolare caratteristica della dialettica online, la citazione, e mostro come questa sia in stretta relazione con segnali sociali, quali l'amicizia tra gli utenti del forum, l'autorità che gli utenti hanno nel forum, e la struttura dell'intera comunità del forum. Questi risultati permettono di migliorare e raccordare comunicazione e socializzazione nel mondo virtuale.

In ultimo, il terzo contributo è uno studio che rivela come informazioni "agnostiche rispetto al contenuto" rispecchino il comportamento degli utenti. In particolare analizzo i troll, utenti che tramite mendacia e manipolazione causano gravi danni alle comunità virtuali. Infatti, i troll usano un linguaggio che ben si nasconde nelle conversazioni che essi portano al parossismo, rendendo difficile per i moderatori raccogliere prove certe che li smascherino. Nonostante ciò, mostro che è possibile individuare le discussioni che saranno colpite dai troll, e le reazioni degli altri utenti ai loro messaggi, tramite una combinazione di informazioni "agnostiche rispetto al contenuto" e lessicali. Questo studio in particolare sottolinea come le informazioni "agnostiche rispetto al contenuto" possano fornire un punto di vista alternativo e complementare al contenuto dei messaggi.

Applicazioni emergenti come Snapchat, Tumblr, e Yik Yak stanno vieppiù abbandonando il paradigma della comunicazione informatica come discussione persistente, lineare, e testuale, preferendo contenuto effimero, destrutturato, e multimediale. Sebbene i risultati presentati si basino principalmente su web forum, l'impianto teorico e metodologico della tesi generalizza a queste nuove piattaforme, e ad una vasta gamma di paradigmi di comunicazione. Questa tesi vuol essere un passo verso una comprensione più approfondita del non detto nell'interazione virtuale, e di come sia possibile superare i suoi limiti.

Contents

List of Figures

List of Tables

to my brother

1
Introduction

How do we interact when we communicate through a computer? A common approach to tackle this question at the core of human-computer interaction is to focus on the nature of the content we share, or the qualities we impart to that content: "what language do we use to convince people we are trustworthy"[1], "how do we present ourselves through avatars"[2], "what pictures resonate with a young audience"[3]. This dissertation, instead, focuses on the *way* we interact: "at what time of the day do we usually update our status", "whose messages do we reply to the most", "do we prefer to engage with many or with few people at a time". In other words, this dissertation looks at the digital traces of the *act* of handling content, disregarding the content involved. I call these digital traces *interaction patterns*.

[1] Soni et al., "Modeling Factuality Judgments in Social Media Text", 2014

[2] Hum et al., "A picture is worth a thousand words: A content analysis of Facebook profile photographs", 2011

[3] Han et al., "" Teens are from Mars , Adults are f rom Venus " : Analyzing and Predicting Age Groups with Behavioral Characteristics in Instagram", 2016

Interaction patterns may appear a rudimentary approach to analyzing online interaction. We often think of interaction patterns as a side effect of a device mediating our communications, instead of a conscious effort to communicate something – we may update our status right after waking up because picking up our smartphone is part of our morning routine, or we may reply to several messages in one go because an interface conveniently allows us to. Also, it is natural to assume that the content we share reveals our intentions more directly than our way of sharing it – after all, it is through that content that we try to communicate.

However, it is easy to see that is not always the case, and much of the communication happens beyond content. When we comment on an old picture we inevitably evoke a sense of nostalgia. When we acknowledge reading a message, and yet decide not to reply to it, we are clearly informing the sender that we are ignoring him. When we forward some content, we may do it to show we endorse that content, or that we support the content's author. Each of these actions carries a rich message, and the content involved is almost irrelevant in understanding the meaning of the message. Following the old adage, *"it isn't what you do, but how you do it"*.

In the offline context, we are accustomed to the idea that appearance and gestures characterize ourselves as individuals, communicate our status, and clarify our role in the context of an interaction, independently from the content we communicate. Nonverbal cues

appear under many names on an extensive track record of research, ranging from biology to social psychology. In the online context, on the contrary, it is common to assume that online communication is natively devoid of nonverbal cues, and that communication applications need to enrich content through interface add-ons – for example emojis, tags, and social buttons.

Nonetheless, the importance of interaction patterns in online communication has been clear from very early on.Pioneering researchers were enthusiastic to observe that the users of Habitat, the first graphic-based massively multiplayer online role-playing game (MMORPG),[4] developed a lingo to convey nonverbal context (referred to with the Japanese term *"kansei"*[5]):

> Yoshida and Kakuta ... specifically compare the human interface in communications technologies to the notion of kansei. ... Joichi Ito ... also emphasizes the need for understanding kansei when evaluating communications in Japan, even online–especially online, where many of the acutely important social cues are missing. ... Kansei might turn out to be an important term all over the Net, as an aid to evaluating the advantages and disadvantages of each Net tool in different situations.[6]

[4] `https://en.wikipedia.org/wiki/Habitat_(video_game)`, accessed 21/1/2017

[5] Kansei is a broader concept than nonverbal cues, that can be loosely translated as *"an intuitive, partially aesthetic sense of rightness about the contextual elements in a conversation."* (Rheingold 1986)

[6] Rheingold, *The Virtual Community*, 1986

In the early days of virtual communities, the ability to convey nonverbal cues appeared an intuitive measure to compare online communication tools in the forthcoming future. However, such a comparison is extremely difficult in practice. We cannot leverage our offline experience as a yardstick, since many elements crucial to our nonverbal vocabulary, such as eye gaze or handshake[7], failed to find their way into consumer products. Moreover, applications differentiate themselves through signature interaction idioms (e.g. like, mention, hashtag), which leave little common ground to compare different applications on. Perhaps as a consequence, research is increasingly skeptical in considering online interaction patterns as universal characteristics of human communication, as it is for their offline counterparts. Instead, online interaction patterns are studied within the realm of individual communities, and often as subordinate information to text and images, that offer clear offline comparisons.

[7] Sumi et al., "Collaborative capturing, interpreting, and sharing of experiences", 2006; Kunii et al., "Telehandshake using HandShake Device", 1995

Despite the relevance of interaction patterns in everyday communication, therefore, there is no general framing for interaction patterns online. This dissertation attempts to fill this gap. I argue that all online communication natively carries interaction patterns, as a natural outspring of the simple actions involved in handling content: e.g. the time a message is shared, and how far into the conversation it appears. With this mindset, I propose a general framing for analyzing online discussion that puts interaction patterns at its core, disregarding the content they refer to. Such a framing highlights a number of possible uses of interaction patterns in enhancing online communication.

Social media are torn between the need to tailor their offer to their users, and the privacy and copyright concerns that come with

learning from the users' content: interaction patterns are direct way to measure user activity in aggregate with minimal disclosure requirements. Also, identity theft and the spread of false information are deeply rooted problems in social media. While deceitful content is by definition difficult to unmask, interaction patterns are more difficult to consciously manipulate, and may therefore be more revealing. Finally, a better understanding of interaction patterns allows existing conversational interfaces to make better use of their expressive potential, and highlights missed interaction opportunities for new interfaces.

We share content in many formats and through many channels, and new ways emerge as technology advances. Still, basic questions on how to make sense of online interaction remain unanswered. If we ignore the content of a discussion, is the way people interact part of their personal style? Does it convey social signals? If so, how could online discussion effectively leverage this information?

1.1 *What are online interaction patterns, exactly?*

This dissertation revolves around interaction patterns in online discussions, defined as the characterization of how users communicate through the platform and with each other, regardless of the communication content. However, it is often difficult to draw the line between what is content and what is not. It is reasonable to consider content the text in a text message, and not to consider content the position of the cell tower that forwards the message. But – recalling our previous examples – is the notification that the receiver opened a message *content*? What about the timestamp of a picture?

A more common term in online communications that comes to mind when thinking of interaction patterns is metadata. Metadata describe data, their structure, and any additional information that can help manage a resource.[8] Like interaction patterns, metadata provide contextual information that is somewhat independent from the content it refers to, and may reflect the intentions of the author of the content.[9] However, unlike interaction patterns, the purpose of metadata is to make tracking and working with specific data easier. On the one hand, metadata mostly refer to static information. On the other hand, the definition of metadata varies across different contexts, as it is strictly tied to the task at hand.

Overall, no existing definitions for interaction patterns in online discussions are entirely satisfying. Existing terms are either fuzzy, not analytical, or too limited. This dissertation attempts to provide a definition overcoming these limitations, building upon the intuitive meaning of interaction pattern and its empirical difference from content.

[8] `https://en.wikipedia.org/wiki/Metadata`, accessed 21/1/2017

[9] The Electronic Frontier Foundation best exemplifies how: *"They know you called the suicide prevention hotline from the Golden Gate Bridge. But the topic of the call remains a secret."* `https://www.eff.org/deeplinks/2013/06/why-metadata-matters`, accessed 21/1/2017

1.2 Scope

This dissertation is a small step towards understanding interaction patterns in online discussion. While they apply to all online communication, for obvious reason this dissertation only covers selected cases. For all analyses I use historical data of public multi-party online discussion – therefore not covering cases of private[10] or ephemeral[11] interaction patterns. Also, analyses often concentrate on case-study interaction patterns. However, these should not be seen as limits of this work. The specific research questions in this work all focus on the feasibility of linking features of discussions to information about users that take part in them. Therefore, the results should be seen as general, proving basic properties of interaction patterns.

[10] e.g. an individual user's browsing history

[11] e.g. the real-time feedback that someone is typing

1.3 Existing approaches

Many investigations of online communities have focused their attention, at least in part, on the interaction patterns of online discussion. Frequent questions that involve interaction patterns are "do people have a recognizable style when they write messages?"[12], "do users that behave similarly show similar patterns in what they share online?"[13], "do successful discussions evolve similarly?"[14]. However, most of this research relegates interaction pattern to the role of supplementary features, and concentrates instead on discussion content (usually, text). This work, instead, makes interaction patterns its primary focus, and considers them informative regardless of the content of the discussion.

[12] Abbasi et al., "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace", 2008

[13] Cheng et al., "Antisocial Behavior in Online Discussion Communities", 2015

[14] Aumayr et al., "Reconstruction of Threaded Conversations in Online Discussion Forums", 2011

Computational models used for research on online communities often incorporate features other than text to boost performance. Depending on the research question and the nature of the data, they may incorporate features on links[15], quotes[16], hashtags, @mentions, retweets[17]. However, are these features only correlates of the main content, or do they carry any information in and of themselves? If they do carry information, how can we interpret its meaning? For example, quotes help identify who is the author of a message in online forums[18]. Is it because quotes are part of our personal writing style, or just because we are interested in different topics that we happen to quote a lot? This dissertation shows that, indeed, quotes are part of our personal writing style. But analyzing quotes in isolation from content tells us more: we use quotes to send social signals – the way two users quote each other tells us if they are friends or not. To obtain these results we must consider content and interaction patterns separately, and understand they respective role and meaning.

[15] De Vel et al., "Mining e-mail content for author identification forensics", 2001

[16] Barcellini et al., "A study of online discussions in an open-source software: Community reconstructing thematic coherence and argumentation from quotation practices", 2005

[17] Arakawa et al., "Adding twitter-specific features to stylistic features for classifying tweets by user type and number of retweets", 2014

[18] Abbasi et al., "Applying authorship analysis to extremist-group web forum messages", 2005

1.4 Contribution

This dissertation makes the following contributions:

1. **An empirical definition of interaction patterns**

   Literature lacks a clear definition of interaction patterns. It has substituted several terms depending on the task at hand, from structural features to interaction cues, that either fail to generalize outside single application domains, or are too diffuse a concept to be practically useful. Chapter 4 gives an operative definition of interaction patterns, that is both general across different forms of online interaction (including e.g. non-textual interaction), and directly actionable for quantitative analysis. Albeit formal, this is a fundamental step in analyzing interaction patterns as a stand-alone element of communication.

2. **A working proof that interaction patterns characterize users' personal contribution styles**

   Chapter 4 proves that interaction patterns can reveal the author of a message, without employing any information about the actual content of the message. Focusing on online forums, it distills a case-study set of interaction pattern features, that completely disregard post content. Then, it proves that these features can accurately predict who is the author of a post, using data from four online forums. I specifically chose four forums with different size, language, and topic, so as to minimize bias deriving from content and scale. Chapter 3 introduces the four forums. A simple classification testbed, relying exclusively on interaction patterns, confirms the author of a message with 76% accuracy, and discriminates between two candidate authors with 94% accuracy. This is the first study to prove that interaction patterns *in and of themselves* carry information about how users interact.

3. **Findings on how interaction patters link discussion to social signals**

   Chapter 5 focuses on one mode of interaction, the quote, and uses it to investigate the structure of the communities in the four forums presented in Chapter 3. Quotes are features of online discussion interfaces that help keep conversation on topic in multiparty discussion[19]. Previous research also associated quotes with signals of acknowledgement, endorsement, and attribution in interpersonal relationships[20]. This work models how users quote each other in discussions, and uses this model to explain characteristics of a forum's community. At a relational level, quoting patterns predict if two users are friends or not with reasonable accuracy. At a community level, quoting patterns reveal users with leading roles, in some cases better than the user profiles themselves do. This proves we can infer social traits of users from the way they interact in discussion, disregarding the explicit social signals typical of modern social networks, such as friends and followers.

4. **Applications of interaction patterns to identifying abusive behaviour**

[19] Barcellini et al., 2005; Li et al., "Modeling Interactions in Web Forums", 2014

[20] boyd et al., "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter", 2010; Garimella et al., "Quantifying Controversy in Social Media", 2015

While Chapters 4 and 5 respectively show that interaction patterns can reveal personal and relational characteristics of users, Chapter 6 demonstrates their use to identify user behaviour. It focuses on *trolling*, a particularly disrupting form of online abuse. In 2015, in an internal memo, Twitter CEO Dick Costolo stated: *"We lose core user after core user by not addressing simple trolling issues that they face every day"*[21]. To this day, despite the apparent simplicity of curtailing abuse, trolls skillfully deceive automated and human moderators. This work gives quantitative insights on why that may be the case, and how to frame them instead. I show that it is difficult to detect trolls from the surrounding discussion, because the content they post finds effective camouflage within a discussion with similarly heated tones . However, interaction patterns detect discussions that will eventually be trolled with high accuracy, and responses to troll posts show consistent patterns in their content. Results show that finding conversation that will be trolled, and tracking down troll posts through their responses, seems a more effective strategy than directly targeting troll posts. This work also shows how interaction patterns provide information that is distinct from – but complementary to – message content. In particular, interaction patterns are more difficult to consciously manipulate than text, and may therefore be more truthful signals of malicious behavior.

[21] Buni et al., *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*, 2016

The ability to enrage others while remaining covert is, after all, the primary characteristic of trolls.

1.5 *Arc of this dissertation*

The rest of this dissertation is organized as follows. While existing theory lacks satisfying definitions of interaction patterns, considerable literature has addressed concepts related to interaction patterns, analyzed specific interaction patterns, or used them in tasks relevant to this dissertation. Chapter 2, provides Chapter 3 describes the four forums used as the source of data for the rest of the dissertation: it clarifies the contents of the dataset, and outlines the advantages and limitations deriving from its use. The following chapters report the results of this line of research. First, Chapter 4 lays the backbone of this work, proposing an actionable, general definition of interaction patterns. Leveraging such definition, it extracts a case-study set of interaction-pattern features for forum posts, and proves that it is possible to identify the author of a post looking solely at interaction patterns, while completely disregarding post content. Chapter 5 then investigates the role of interaction patterns beyond the individual. Discussion forums are online communities – however, they lack (or see minimal use of) the explicit social signals we grew accustomed to in social networks, such as befriending, following, or reputation. Chapter 5 investigates the links between interaction patterns in discussion, and the communities in the forums. Results show that quotes reveal friendship ties between users, and that they mirror characteristics of the under-

lying social structure in the forums. Chapter 6 demonstrates how interaction patterns may complement content analysis in explaining user behavior. In fact, trolls, a deceptive kind of online abusers, elude identification through content alone. A combination of content analysis and interaction patterns, however, exposes them, and may greatly help moderators. Finally, Chapter 7 discusses the results presented in this dissertation, and its limitations, and outlines how it may inform future research.

2
Related work

This work uses interaction patterns as the unit of analysis of online activity: it wouldn't be feasible reinterpret all work on social media and human-computer interaction through the lens of interaction patterns. This chapter focuses solely on the core concepts relevant to the topic at hand, and reports the corresponding seminal findings, as well as those offering the most interesting prospects of future research. For the sake of readability, I postpone presenting further work that is essential to interpreting results, but does not add to the state of the art on interaction patterns, until the corresponding results are presented, in subsequent chapters.

Although there is a truly vast literature on computer-mediated communication making at least marginal use of interaction patterns, only a fraction of it has interaction patterns as its main focus. Instead, interaction patterns are more often tools to measure the context they appear in. Therefore, to make comparison easier, I divide literature according to its scope of analysis: how interaction patterns impact discussion, how they characterize users, how they distinguish relationships between users, and how they reflect the structure of a community. At the end of each section, I highlight how the contributions in this dissertation relate to existing research.

2.1 Interaction in online discussion

Before focusing on what online discussion discloses about users and their relationships, this section presents literature that frames the problem in the reverse direction – how user interaction shapes discussion. I will first review literature that addresses how discussion shapes and evolves in the offline and online domains, before moving on to how specific interactions affect discussion content.

Some of the research in this line of work investigates how groups advance face-to-face conversation through subsequent communication phases – e.g. disclosing information or converging to a decision. Computational models include the detection, discovery, and recognition of which nonverbal cues signal progression in the conversation, through the analysis of transcripts or audio-video recordings[1]. This research shows that nonverbal cues add predictive power to lexical features in inferring who is the current

[1] Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review", 2009

speaker[2], whom the current speaker is addressing[3], who will speak next, and who is present in the conversation[4]. These works suggest that the history of participants' co-presence and turn-taking, among other nonverbal cues, help shape the evolution of a conversation.

Although the methods used for face-to-face discussion analysis are likely not helpful in the online context, where the number of participants is far greater and the raw data far noisier[5], studies on Web and social media also suggest that user interaction affects discussion structure.

Backstrom et al.[6] and Kumar et al.[7] analyze separate large datasets (respectively, Facebook and Wikipedia, and Tiwtter, Yahoo! groups, and Usenet), but share similar findings on how timing and the identities of the participants relate to the structure of a discussion. Backstrom et al. show that threads exhibit a bimodal distribution of the number of participants: they are either dominated by a very small number of distinct users, or by many users who generally post only once. Kumar et al. further show that a branching model is able to cluster dyadic and group discussions. Moreover, Backstrom et al. show that threads are significantly longer in Facebook when the first replies come from friends, and when the first replies arrive early. Kumar et al. further show that a preferential attachment generative model that accounts for recency explains well the reply structure of a thread. Also, Backstrom et al. show that patterns of appearance of first commenters in the thread are predictive of whether the user that started the thread will comment again. Kumar et al. shows that a Polya urn process that accounts for authors responding to responses to their own earlier messages explains the arrival patterns well. Aumayr et al.[8] expand analyses reconstructing which posts reply to which others in a thread. They use a larger set of features, comprising many nonverbal including timing, quotes, post index, and thread length, which show the best precision and F1 score[9].

Besides discussion structure and evolution, several works investigate how specific types user interactions affect discussion content. Particularly relevant to this dissertation is work focusing on quotes. Quotes signal shared attention and addressee acknowledgement in discussion. Literature shows that quotes help highlight the focal points in a discussion, and maintain the discussion on topic[10]. On Twitter, quote-retweets seem to encourage longer and more civil discussion[11]. Recent research has built tools to interpret public dialogue through quoted text, which can expose the systematic bias in news media outlets[12].

[2] Canseco et al., "A comparative study using manual and automatic transcriptions for diarization", 2005

[3] Akker et al., "A comparison of addressee detection methods for multiparty conversations", 2009

[4] Chaudhuri et al., "A comparison of latent variable models for conversation analysis", 2011

[5] Shriberg, "Spontaneous speech: How people really talk and why engineers should care", 2005

[6] Backstrom et al., "Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry", 2013

[7] Kumar et al., "Dynamics of Conversations", 2010

[8] Aumayr et al., "Reconstruction of Threaded Conversations in Online Discussion Forums", 2011

[9] F1 is a measure of prediction accuracy that balances precision and recall:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

[10] Barcellini et al., "A socio-cognitive analysis of online design discussions in an Open Source Software community", 2008; Kang et al., "Analyzing answers in threaded discussions using a role-based information network", 2011

[11] Garimella et al., "Quote RTs on Twitter", 2016

[12] Niculae et al., "QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns", 2015

This work

Literature shows that the way users interact affects the structure, content, and evolution of a discussion. However, the focus of work in this area is typically on the discussion itself, rather than on how users interact; this dissertation on the other hand investigates to

what extent interaction patterns in online discussion provide information on the participants. Nonetheless, the results from the two lines of research show promising correlations. In fact, the results presented above inform the choice the interaction pattern features for forum posts presented in Chapter 4.

Chapter 5, although focused on reconstructing social structure from quotes, presents novel findings on how quotes shape discussion: it adds to current knowledge on thread structure by showing novel relationships between the distributions of posts, threads, and users, and extends previous findings on interaction patterns explaining *how* quotes may support longer discussion and maintain thematic coherence (by relaying context between posts that are far apart in time, and by helping shorten discussion efficiently).

2.2 *Interaction as self expression*

A second body of literature, thematically closer to this dissertation, investigates what we can learn about the individual user from how he interacts. This section mainly focuses on work on detection of an author's style, but also provides pointers to relevant work on privacy and security in online social networks.

Identifying the creator of a portion of content is a task of great interest, both theoretical and practical[13]. Authorship analysis is a long-standing field of research that associates written material to author profiles, based on the idea that authors have a persistent and unique writing style subconsciously imparted to their entire production[14]. Applications in the online domain include digital humanities, user profiling, and digital forensics.

Authorship analysis typically addresses three major tasks[15]:

Authorship attribution: Identify the author of an anonymous text among a predefined set of candidate authors, comparing the anonymous text to texts indisputably written by the candidates. This task is often modeled as a multiclass, single label classification problem, where the input are the features extracted from each document, and the output label is the identity of the most likely author.

Authorship characterization: Infer some profiling information on the authors of anonymous text, other than their identity. The target characteristics may be extremely varied – from the author's gender[16], to his native language[17] and personality traits[18]. This problem may be modeled in a way similar to authorship attribution, where the outcome variable is the target characteristic – depending on the number and nature of the characteristics, it is conceptually easy to adapt it from single- to multilabel, and from categorical to ordinal or continuous output.

Authorship verification: Given two texts, decide if they have been written by the same author, without necessarily inferring the

[13] A natural, and more general, question would be *"what do we communicate through online interaction patterns?"*. This dissertation does not investigate what is the *meaning* of interaction patterns, but only if they are informative at all, which is a more basic question and one more amenable to quantitative analysis. For an interesting theory on the meaning of online interaction patterns, see Donath, "Signals , cues and meaning", 2011

[14] Rudman, "The State of Non-Traditional Authorship Attribution Studies—2012: Some Problems and Solutions", 2012

[15] Zheng et al., "A framework for authorship identification of online messages: Writing-style features and classification techniques", 2006

[16] Koppel, "Automatically Categorizing Written Texts by Author Gender", 2002

[17] Koppel et al., "Computational Methods in Authorship Attribution", 2008

[18] Noecker et al., "Psychological profiling through textual analysis", 2013

identity of that author. This task is sometimes referred to as *similarity detection*. In theory, this task is best modeled as an outlier detection (or one-class classification) problem, where the training set only holds text by the primary author, and "impostor" texts are detected as anomalies in a semi- or unsupervised fashion[19]. However, since it is possible to gather large outlier samples, and supervised algorithms are generally more accurate than semi-supervised ones, this task is implemented in practice as a binary classification problem, where the negative class is a collection of texts by the impostors[20].

[19] Koppel et al., "Measuring differentiability: unmasking pseudonymous authors", 2007

[20] Brocardo et al., "Authorship verification of e-mail and tweet messages applied for continuous authentication", 2014; Koppel et al., "The "Fundamental Problem" of Authorship Attribution", 2012

Over time, authorship analysis literature proposed a large body of features, to boost prediction accuracy[21]. However, it is not clear which are the best features, or even the best feature types, as this may depend on the application[22]. A taxonomy of features, and a rationale behind their use, is the following – which provides motivation for the introduction of interaction-pattern features, and a baseline to evaluate their information content.

[21] Abbasi et al., "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace", 2008

[22] Stamatatos, "A survey of modern authorship attribution methods", 2009

Character features provide basic, character-level statistics of writing style, such as letter count, character type frequency (upper/lower case, alphabetic/digit, punctuation mark), character *n*-grams, or analysis of character sequence repetitions via byte-level compression. While most character features do not require specialized tools for extraction, and prove robust across different languages[23], these features often cannot capture subtle aspects of an author's style.

[23] Peng et al., "Language independent authorship attribution using character level language models", 2003

Lexical features consider text as a sequence of tokens (words, numbers, and punctuation marks). Common features include word and sentence length, vocabulary richness, word frequencies, word *n*-grams, and writing errors. Lexical features give a simple and natural representation of text[24]. However, tokenization is not a trivial task in languages like Chinese[25], and there is no consensus on the extraction procedure, e.g. which (and how many) frequent words to consider.

[24] Burrows, "'Delta': a measure of stylistic difference and a guide to likely authorship", 2002

[25] Li et al., "From fingerprint to writeprint", 2006

Syntactic features leverage authors' unconscious use of similar sentence structures[26], and are therefore considered more reliable than lexical features[27]. On the other hand, these features require robust, accurate, and language-dependent natural-language-processing tools.

[26] Pillay et al., "Authorship attribution of web forum posts", 2010

[27] Stamatatos et al., "Computer-based Authorship Attribution without Lexical Measures", 2001

Semantic features involve higher-level interpretation of content. Some works include features such as word synonyms, semantic dependencies [28], topics[29], emotions[30], and perception[31]. However, semantic features suffer from the same drawbacks as syntactic features, depending on sophisticated semantic taggers in addition to the above processing tools.

[28] Zhang et al., "Authorship identification from unstructured texts", 2014

[29] Seroussi et al., "Authorship Attribution with Latent Dirichlet Allocation", 2011

[30] Mohtasseb et al., "More blogging features for author identification", 2009

[31] Bogdanova et al., "Cross-Language Authorship Attribution", 2014

Application-specific features , finally, exploit characteristics of the given text domain (e.g. electronic mail or microblogging mes-

sages), or specific to the text's language (e.g. diacritics). They can be further divided into *content-dependent* features, that consider contextualized text content (e.g. detecting keywords like *'sale'* or *'obo'* in classified ads), and *structural* features, that consider user habits beyond writing content (e.g. considering layout, formatting, links, use of quotes, greetings, signatures[32], font styles[33], hashtags, @mentions and retweets[34]).

Authorship analysis historically targets handwritten prose of single authors. It comes as no surprise that the online context challenges its traditional approaches[35]: online text is often short, misshapen, or multilingual, which makes extraction of most non-simplistic textual features inaccurate or even impossible[36]. Structural features, the closest feature set to interaction patterns, are to some extent decoupled from text, and have therefore seen increasing use[37]. However, authorship analysis literature lacks a clear, general, and operational definition that clarifies their dependency on content. Also, structural features are typically added, in an *ad-hoc* fashion, to classifiers based mostly on other features, which makes it hard to understand the amount of information they provide.

Another research area that tackles identification of users, on the basis of their social graph, is network security. Narayanan et al.[38] rely on network topology to unmask nodes in an anonymous social network; however, network topology is rarely available to the general public. Govindan et al.[39] restrict the necessary background knowledge to topological features of nodes and nodes in their ego-network; however, the proposed algorithm outputs a set of candidates, and its performance metric is relative to the number of nodes in the network, which makes it difficult to compare results to authorship attribution. Koessler Gosnell[40] uses information about local interaction to unmask nodes; however, its preliminary results are validated on synthetic data only.

[32] De Vel et al., "Mining e-mail content for author identification forensics", 2001

[33] Abbasi et al., "Applying authorship analysis to extremist-group web forum messages", 2005

[34] Arakawa et al., "Adding twitter-specific features to stylistic features for classifying tweets by user type and number of retweets", 2014

[35] Koppel et al., "Authorship Attribution: What's Easy and What's Hard?", 2013

[36] De Vel, "Mining e-mail authorship", 2000; Eder, "Does size matter? Authorship attribution, small samples, big problem", 2014; Juola, "Future trends in authorship attribution", 2007

[37] Juola, "Authorship Attribution", 2007

[38] Narayanan et al., "De-anonymizing social networks", 2009

[39] Govindan et al., "Local Structural Features Threaten Privacy across Social Networks", 2013

[40] Koessler Gosnell, "Social Fingerprinting : Identifying Users of Social Networks by their Data Footprint", 2014

This work

Authorship analysis literature conflates content and interaction patterns, partly because of its heritage of text analysis, partly for the lack of a formal distinction between these different sources of information. This work, instead, shows that interaction patterns by themselves provide information on users. Nonetheless, the typical framing of authorship analysis proposes is crucial to this dissertation. Chapters 4 and 5 adopt its formalization of attributing posts to users as authorship attribution and verification tasks (this work does not tackle authorship characterization). Also, Chapters 4, 5 and 6 employ its modeling of attribution as classification problems. Moreover, the definition of interaction patterns as content-agnostic features in Chapter 4 finds its closest match in existing literature in the concept of structural features (albeit "structural" is a *de-facto* moniker, rather than a well defined category). Chapter 4 shows that a few content-agnostic features yield state-of-the-art performance

on forum posts, comparable to lexical features. Chapter 6 provides an empirical demonstration that semantic and content-agnostic features have additive predictive power that reflects different aspects of user behavior. Chapter 5 performs network de-anonymization, through user interaction – while most literature in the field leverages instead social network edge information.

2.3 *Repeated interaction as social signal*

Social media allow users to maintain important relationships. However, the converse is not true: not all links in social networks correspond to relationships that users find important. What does it mean to be friends on Facebook[41]? A large body of literature links the way pairs of users interact, and the real-life meaning of their relationship. First, I review literature that tries to explain interaction between existing online relationships. Then, I summarize research that focuses on the act of creating a new online relationship.

Several social science theories customarily support research in this direction. A simple principle governing user interaction is homophily: users with similar characteristics are more likely to establish relationships. Online media show evidence that this phenomenon also drives content consumption: online friends consume similar content [42,43]. Literature on tie strength posits that not all relationships are created equal: for example, we have strong ties with very good friends, and weak ties with acquaintances[44]. Gilbert et al.[45] present an analytical framework to compute tie strength on Facebook. They draw from social science theory to craft meaningful features of user profiles, interaction, content, and social network structure. The resulting model differentiates gold-standard strong and weak ties with high accuracy. Intimate interaction, together with high interaction intensity, are the feature categories that best predict tie strength. The best individual features are the timespan of the interaction history, and the recency of the last interaction – network structure alone is a weak predictor, but becomes the third-most powerful in interaction with other dimensions. In later work, Gilbert[46] ports the same model to Twitter, essentially confirming previous results.

A large body of work studies characteristics of interaction to infer characteristics of relationship, and vice versa[47], e.g. analyzing professional[48] or romantic status[49]. Interestingly, Burke et al.[50] find that directed communication is associated with greater feelings of bonding social capital and lower loneliness.

A different body of literature focuses on the dynamics of the formation of new online relationships from existing ones[51]. A theory that supposedly drives relationship building is social balance, or triadic closure – in brief, common friends are more likely to create friendship[52]. A simple model, where one user chooses an existing friend at random, and befriends one of his friend's existing friends chosen at random (among the ones the user is not already friends

[41] Wilson et al., "User interactions in social networks and their implications", 2009

[42] Chang et al., "Specialization, Homophily, and Gender in a Social Curation Site: Findings from Pinterest", 2014;Aiello et al., "Friendship prediction and homophily in social media", 2012

[43] Albeit at the cost of limiting exposure to diverse information: see Graells-Garrido et al., "Data Portraits and Intermediary Topics: Encouraging Exploration of Politically Diverse Profiles", 2016

[44] Granovetter, "The strength of weak ties: A network theory revisited", 1983

[45] Gilbert et al., "Predicting tie strength with social media", 2009

[46] Gilbert, "Predicting tie strength in a new medium", 2012

[47] Wilson et al., "A Review of Facebook Research in the Social Sciences", 2012

[48] Dino et al., *Online Interactions Between Group Members Who Differ in Status*, 2008;Owens et al., "Technologies of Status Negotiation: Status Dynamics in Email Discussion Groups", 2000;Mitra et al., "Analyzing Gossip in Workplace Email", 2013

[49] Backstrom et al., "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook", 2014

[50] Burke et al., "Social Network Activity and Social Well-Being", 2010

[51] Liben-Nowell et al., "The link-prediction problem for social networks", 2007

[52] Hutto et al., "A longitudinal study of follow predictors on twitter", 2013

with), fits social media data better than preferential attachment – a.k.a. rich-get-richer, where more visible users are more likely to attract more friends[53]. The microscopic operation of adding an edge can explain the macroscopic evolution of the social graph[54]. One related field of research considers links as representing interaction, instead of relationship, and analyzes content diffusion. Applications range from diffusion of retweets[55] to rumors[56] to memes[57] to emotion[58]. Only very recent work however addresses how interaction affects the creation of a new relationship edge[59].

[53] Leskovec et al., *Mining of massive datasets*, 2014

[54] Leskovec et al., "Graphs over time: Densification Laws, Shrinking Diameters and Possible Explanations", 2005; Aggarwal et al., "Evolutionary network analysis: A survey", 2014

[55] Kwak et al., "What is Twitter, a social network or a news media?", 2010

[56] Friggeri et al., "Rumor Cascades", 2014

[57] Leskovec et al., "Meme-tracking and the Dynamics of the News Cycle", 2009

[58] Kramer et al., "Experimental evidence of massivescale emotional contagion through social networks", 2014; Ferrara et al., "Measuring emotional contagion in social media", 2015

[59] Farajtabar et al., "COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution", 2015

This work

Literature on tie strength, and more in general literature that characterizes relationships through online interaction, assumes that a link between the two users exists. Chapter 5 addresses predicting the existence of such a link using interaction patterns – specifically, quotes. Aiello et al.[60] perform a similar task; however they ground their analyses in the similarity of the content two users consume, instead of the interaction between users. The formulation of the friendship prediction task in this dissertation also differs from that of most literature: link prediction usually infers a new link within a social graph; this dissertation instead predicts a new link in the social graph from a (distinct) interaction graph. Wilson et al.[61] formulate the problem in a similar fashion; however, they do so on Facebook, where users interact with existing friends: their interaction graph is an overlay of the underlying social graph, while in this dissertation the two graphs are (surprisingly) distinct. This is a relatively novel approach, and the fact that most real-life social graph information is not public makes it all the more valuable.

[60] Aiello et al., 2012

[61] Wilson et al., 2009

Chapter 5 confirms the findings of high triadic closure and no rich-get-richer in quoting interactions, as suggested by the generative models for social networks in Leskovec et al.[62]. Chapter 4 and 5 draw inspiration in the choice of features for authorship analysis, deanonymization, and friendship prediction from literature on tie strength. Chapter 4 in particular draws ispiration from methods and feature grouping proposed by Gilbert et al.

[62] Leskovec et al., 2014

2.4 *Collective interaction as community structure*

This section reviews literature that uses interaction patterns to understand the composition of online communities. Research in the field faces limitations similar to those examined in the previous section: the social graph is often unavailable, and even when available, online friendship links are often not meaningful. As a consequence, this dissertation ignores the explicit social network, where users explicitly signal their friends and followers, and concentrates on the network of interactions between users. This is often referred to as an implicit social network, although it is important to note than in our case links are interactions instead of social links.

There is substantial evidence that online interaction graphs share properties of social graphs[63]: it seems that all human social behaviours share some universal (not yet fully understood) patterns[64]. As a proxy for social signals, research substitutes interaction such as co-presence at events [65], academic citations [66], emails [67], phone calls [68], private messages [69] and replies in online discussion[70]. Most previous work builds the implicit social graph as the graph where users are nodes, and the edge between two users is weighted by the number of interactions between them – possibly eliding edges with weight smaller than a threshold, and/or binarizing them (nodes are either connected with equal weights, or not connected). Gupte et al. propose an axiomatic way to construct an implicit social network from desired properties of tie strength. This *de-facto* structure of the community can then be used to answer questions typical of social networks: predicting node popularity[71], finding influential nodes[72], characterizing user roles[73] and reputation[74].

De Choudhury et al.[75] warn that different ways of defining social ties from interaction (e.g. two users are connected if they exchange at least X emails) result in structurally different implicit networks. Thus, although the implicit social graph is of great interest for its applicability, we cannot assume its structure reflects that of the community it attempts to measure – at least without validation. In particular, few works investigate the relationship between the implicit and the explicit social graph. Zhou et al.[76] propose a theoretical model to overlay different implicit networks that reflect distinct interests in the community. Wilson et al.[77] overlay the implicit and explicit graphs on Facebook, and highlight that few relationships are maintained through interaction. Frey et al.[78] overlay the implicit graph, that connects users based on shared interests, and the explicit graph, that connects users based on trust, and proposes this combination as a platform for trusted transactions. The above research on overlays, however, assumes that interaction happens only between friends in the explicit graph: this is not the case for venues for open discussion, such as online forums, news media sites, or Twitter.

[63] Leskovec et al., "Planetary-scale views on a large instant-messaging network", 2008; Aiello et al., 2012. As a reference, network properties for popular social networks can be found in e.g. Myers et al., "Information network or social network? The Structure of the Twitter Follow Graph", 2014; Ferrara, "A large-scale community structure analysis in Facebook", 2012; Mislove et al., "Measurement and Analysis of Online Social Networks", 2007. Meusel et al., "Graph structure in the web — revisited", 2014 reports network properties for the WWW graph

[64] Barabási, "The origin of bursts and heavy tails in human dynamics", 2005

[65] Zhou et al., "A social network matrix for implicit and explicit social network plates", 2014

[66] Leskovec et al., 2005

[67] Roth et al., "Suggesting Friends Using the Implicit Social Graph", 2010

[68] Gupte et al., "Measuring tie strength in implicit social networks", 2012

[69] Panzarasa et al., "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community", 2009

[70] Gómez et al., "Statistical analysis of the social network and discussion threads in slashdot", 2008; Anwar et al., "Modeling a web forum ecosystem into an enriched social graph", 2013

[71] Hutto et al., 2013

[72] Kempe et al., "Maximizing the spread of influence through a social network", 2003; Shafiq et al., "Identifying leaders and followers in online social networks", 2013

[73] Welser et al., "Visualizing the signatures of social roles in online discussion groups", 2007

[74] Anderson et al., "Discovering value from community activity on focused question answering sites: a case study of stack overflow", 2012

[75] De Choudhury et al., "Inferring relevant social networks from interpersonal communication", 2010

[76] Zhou et al., 2014

[77] Wilson et al., 2009

[78] Frey et al., "Social market: Combining explicit and implicit social networks", 2011

This work

Chapter 5 builds the implicit social network of user quotes. It shows that this network is similar in four different forums, and in all four cases exhibits a social-like structure. While this is in line with previous literature, it is novel in demonstrating that quoting structure is consistent across discussion platforms. Chapter 5 also employs the quote network to reveal influential users in the forums. Unlike previous work, it uses properties of the quote networks to compare *different* communities – in particular, it investigates evidence of power differentials in the user base. Few users in the forums under study use the forums' friendship system – too few

to provide ground truth on the structure of the underlying community. Following De Choudhury et al.'s disclaimer[79], the chapter does assume the quote network replicates the structure of the forums' community. However, it proves a useful tool in inferring properties of the underlying community, such as relationships between users and user roles. This opens promising applications on retrofitting discussion-based communities with social features. Chapter 5 draws from the related literature to analyze the quote and friendship graphs, leveraging some of its insights to select features for friendship prediction.

[79] De Choudhury et al., 2010

This chapter compares and contrasts this dissertation with related literature. In particular, it shows that literature still lacks an actionable definition of interaction patterns in online discussion, and that so far it has mostly used interaction patterns in conjunction with features that depend on message content – it is yet unclear if interaction patterns carry any information on users *per se*. However, there is promising evidence of the converse: users and their ties shape online discussion. Literature shows skepticism on the meaningfulness of friendship links in online social networks, and suggests that substituting friendship for interaction between users may yield a more truthful representation an online community. Nonetheless, it is yet to be proved whether interaction patterns (disregarding content) may directly or indirectly measure relationships between users.

Several research questions and contributions in this dissertation are novel. However, it must be acknowledged that it owes much to previous literature: it borrows framing and modeling from authorship analysis, feature engineering from discourse analysis and tie strength, and analytical methods from graph theory.

This concludes the literature review. The next chapter completes the necessary context for interpreting the analyses, describing the forums that are subject of this study, detailing the crawling process, and giving a quantitative depiction of the data.

3
Data

In many cases, having the right data is more important than elegant theories and sophisticated methods of analysis[1]. This dissertation uses four online discussion forums as the source of data. Forums nowadays are not a "hip" venue for social computing research: unlike several modern social networks, their scale does not even begin to approach that of humanity, nor do they sport the explicit social signals, such as friendship, trust, and group membership, that the field researches or employs as units of measure. However, forum data comes with several advantages that make them a superior choice to properly address the research questions in this dissertation. This chapter explains what these advantages are. Then, it describes the four forums subject of this study, to help get a sense of what the data captures. Finally, this chapter details the data gathering process, provides a quantitative overview of the data, and discusses its limits. But, first, a brief introduction to how forums work is in order.

[1] Halevy et al., "The Unreasonable Effectiveness of Data", 2009

3.1 Interaction in online forums

Forums are a public, online venue for discussion on a topic. Discussions are organized into sections of the forum, called subforums, that address specific aspects of the general topic – e.g. discussion in a music forum may be divided by genre. Individual discussions, called threads, are composed of messages, called posts. Posts are usually mostly text, but may embed emoticons, links, pictures, and videos. Threads start with an opening post (OP) that sets the title and argument of the conversation. Subsequent posts reply to the OP, or to later posts.

Although the abstract data structure for a thread is a tree of replies, where each post comes as a reply to exactly one preceding post, most interfaces show only the linear sequence of the posts, indexed by time of arrival. This is typically broken up over several pages, with each page of the thread showing a window of few to few tens of posts. Users may explicitly refer to previous messages through quotes: users cite excerpts from one or more previous posts, and incorporate them in their message through some code that links back to the original posts (Chapter 5 gives a more thor-

ough description of quotes).

Users may also communicate with each other outside of threads, through private messages. Most forums require registration, and users must log in before posting. Users in the forum know each other through their pseudonymous profile, which in its basic form is a screen name and an avatar. In addition, profiles often show a user's status (for example, if the user is part of the forum staff) and rank (usually a representation of the number of posts they have contributed to the forum, separating newcomers and seasoned members). Some forums include a barebones friendship system: users may, by mutual choice, be listed as friends on each other's profile. However, users do not receive any additional feedback on their friends' activity.

3.2 *Online forums as data sources*

There are three main reasons for choosing forums over other online interaction platforms. First, forums are *transparent*. They are public, and anyone can gain access to them through a simple registration form. Discussion is rarely altered (e.g. split, moved, deleted) over the lifespan of the platform. Almost all discussion is observable – private interaction makes use of different channels and represents only a negligible fraction of the total information exchanged. We see what the users saw when they entered a discussion. This is not the case for several new media, where data presentation is private, personalized, or time-varying. Lack of transparency may result in unrecoverable bias in both how[2] and what[3] we sample, as well as in the user behavior under study[4].

Second, *forum users agree on interaction norms*. Forum conversation primitives are `thread start`, `reply` and `quote`, and their use and meaning have long become unambiguous. The same cannot be said for e.g. Facebook, where users use tags in their posts sometimes to signal the presence of a friend in a past event, and sometimes to attract a friend's attention to an ongoing discussion. For example Garimella et al.[5], in discussing the adoption of quote RTs, rightly warn that *"the usage of this new feature might not have 'converged' yet"*.

Finally, forum discussion is *rich in both content and interaction patterns*, and therefore amenable to analysis through both content-agnostic and content-aware approaches – representing an ideal testbed for comparing the two. For instance, stylometry has been applied successfully to online forums in the past[6], and has showed promising results that suggest that approaches based, respectively, on interaction patterns and on text have complementary strengths.

Discussion forums are a great source of data for many a research question on online behavior. For this dissertation in particular, they allow unobtrusive observation of discussion from the point of view of its participants: this gives a natural representation of users' interaction patterns.

[2] Achlioptas et al., "On the bias of traceroute sampling", 2009

[3] Tufekci, "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls", 2014

[4] Epstein et al., "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections", 2015; Kramer et al., "Experimental evidence of massivescale emotional contagion through social networks", 2014

[5] Garimella et al., "Quantifying Controversy in Social Media", 2015

[6] Abbasi et al., "Applying authorship analysis to extremist-group web forum messages", 2005; Pillay et al., "Authorship attribution of web forum posts", 2010; Zheng et al., "A framework for authorship identification of online messages: Writing-style features and classification techniques", 2006

3.3 The four forums

I chose four forums sporting wide diversity in terms of topic, scale, user background, language, and other factors specific to each community (see Table 3.1 for the number of posts, threads, and users in each forum). On the other hand, I made sure to choose forums employing (customized versions of) the same front-end (vBulletin[7], one of the leading platforms for community software), so as to exclude that any observed differences might stem from the user interface rather than from the forums' intrinsic characteristics. Discussions vary considerably across forums – and indeed even within each forum. For instance, some sections of a forum are dedicated to Q/A, others to review and commenting, and others still to conversation between peers. Figures 3.1,3.2,3.3, and 3.4 show the front pages of the forums, to help get a sense of their content and feel, and brief description of each is provided below.

[7] https://www.vbulletin.com/

RPG is the largest international online forum devoted to roleplaying games (*RPGs*), with a focus on tabletop RPGs. Its users come from many different backgrounds, and include a sizeable minority of professional game developers. The forum is divided into subforums that span a wide range of RPG-related topics, from speculations on new releases to *play-by-post* online games.[8]

[8] See Figure 3.1, http://forum.rpg.net

SWZ is the forum section of an Italian IT news and information website. It serves as a place for knowledge exchange between IT experts and the general public, and its threads feature user-contributed guides, problem troubleshooting, and software/hardware reviews.[9]

[9] See Figure 3.2, http://forum.swzone.it

TM is a major Italian board for discussing metal and hard rock music. Beside areas for casual conversation and music-related classified ads, most conversation revolves around critique of artists and albums, organized in subforums that reflect a taxonomy of subgenres. The community is active and engaged, and encourages users to meet in real life at concerts.[10]

[10] See Figure 3.3, http://truemetal.it/forum

PSY is a mental health support community. It provides information on psychology and personal development. Conversation usually happens in the form of comments either to articles on specific conditions, or to personal stories. The forum, in English, is heavily moderated.[11]

[11] See Figure 3.4, http://forum.psychlinks.ca

3.4 Crawl process and data format

I crawled the four forums, acquiring all posts available from each forum's inception until the day of the crawl. I developed a python script to simulate what a freshly registered user would see logging into the forum, processing the current page top-to-bottom, and browsing to the next. The crawler proceeded breadth-first through the forum structure, first analysing subforums and saving links to

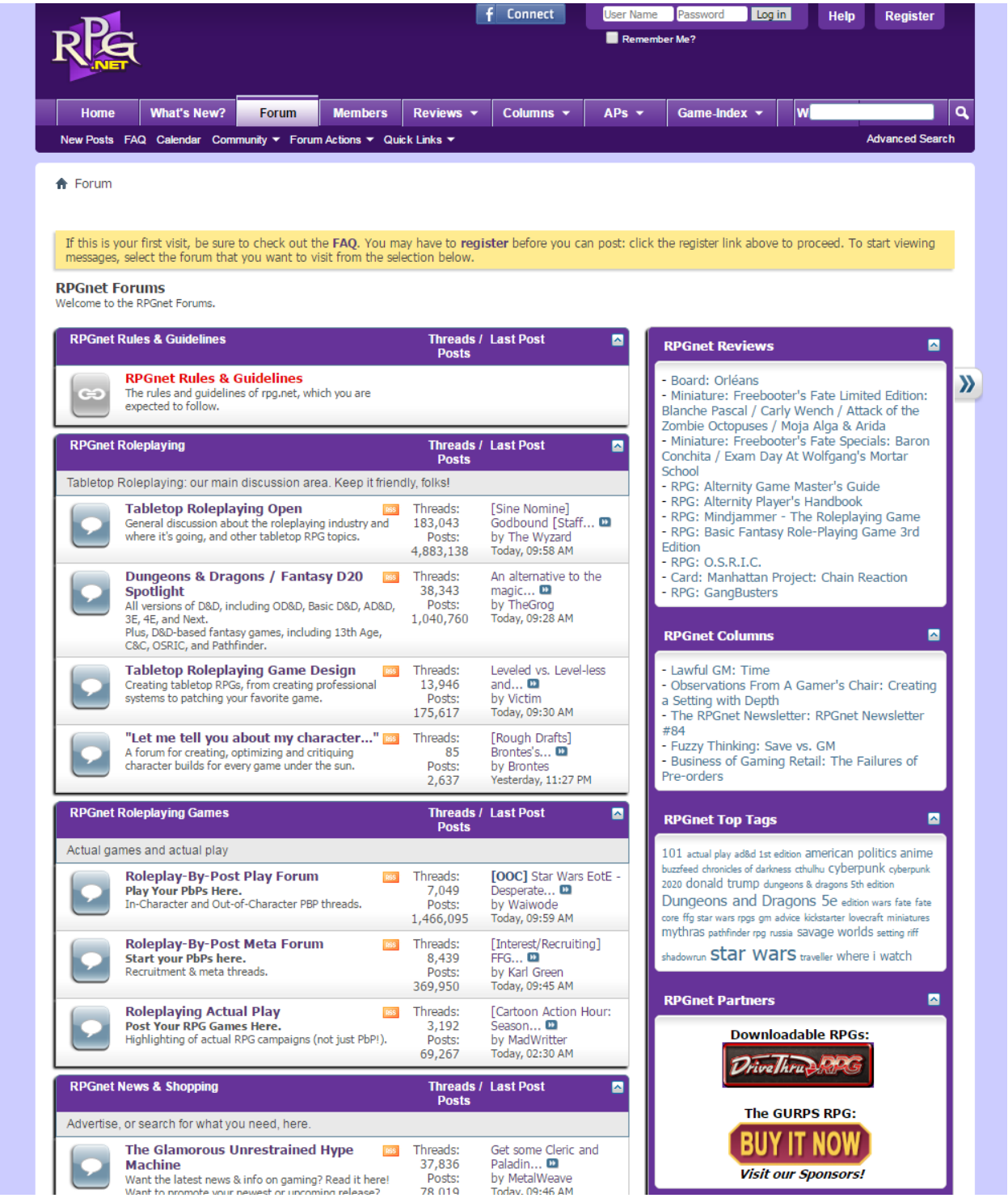


Figure 3.1: Screenshot of the front page of the RPG forum, retrieved on 25/01/17

Figure 3.2: Screenshot of the front page of the SWZ forum, retrieved on 25/01/17

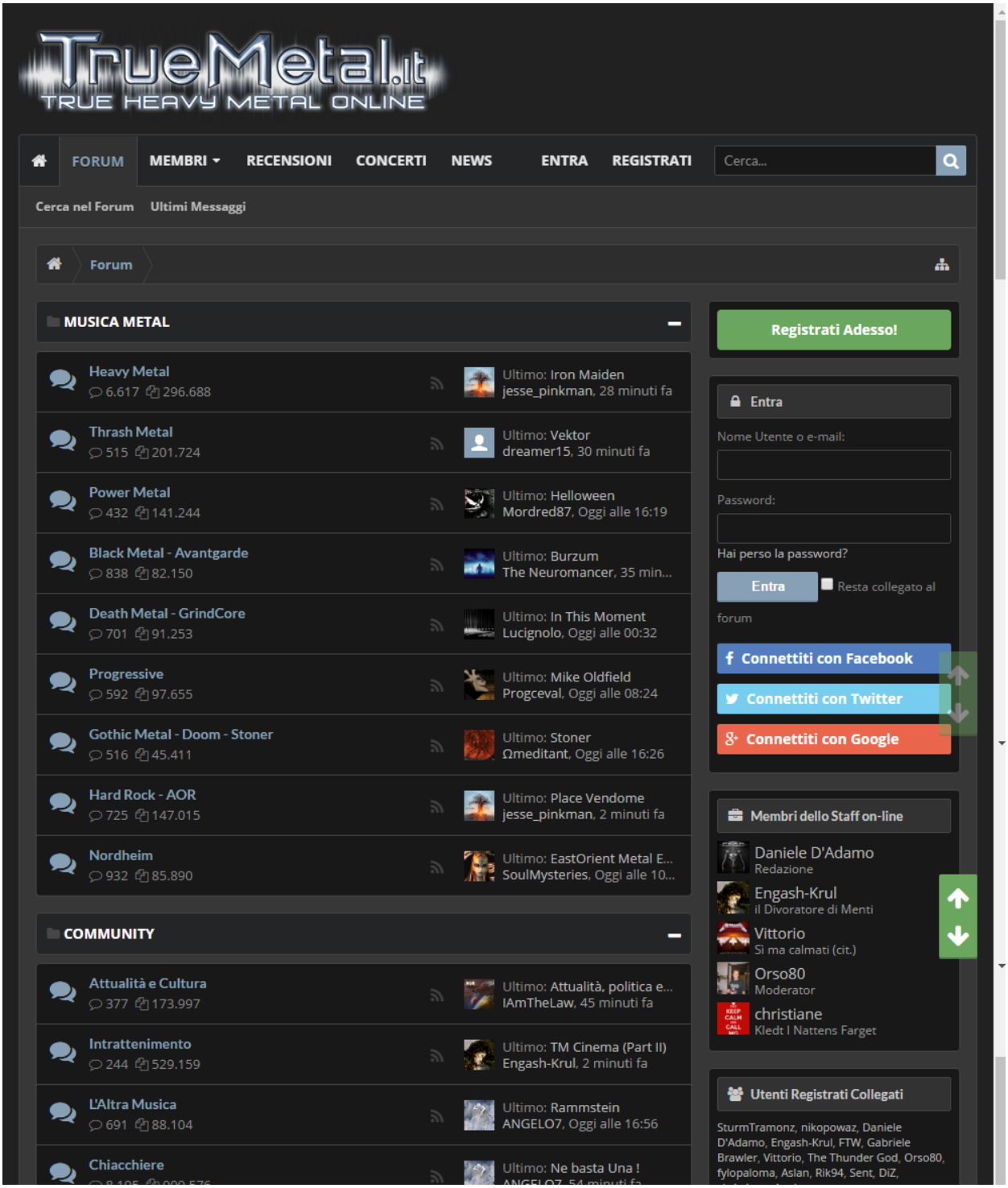


Figure 3.3: Screenshot of the front page of the TM forum, retrieved on 25/01/17

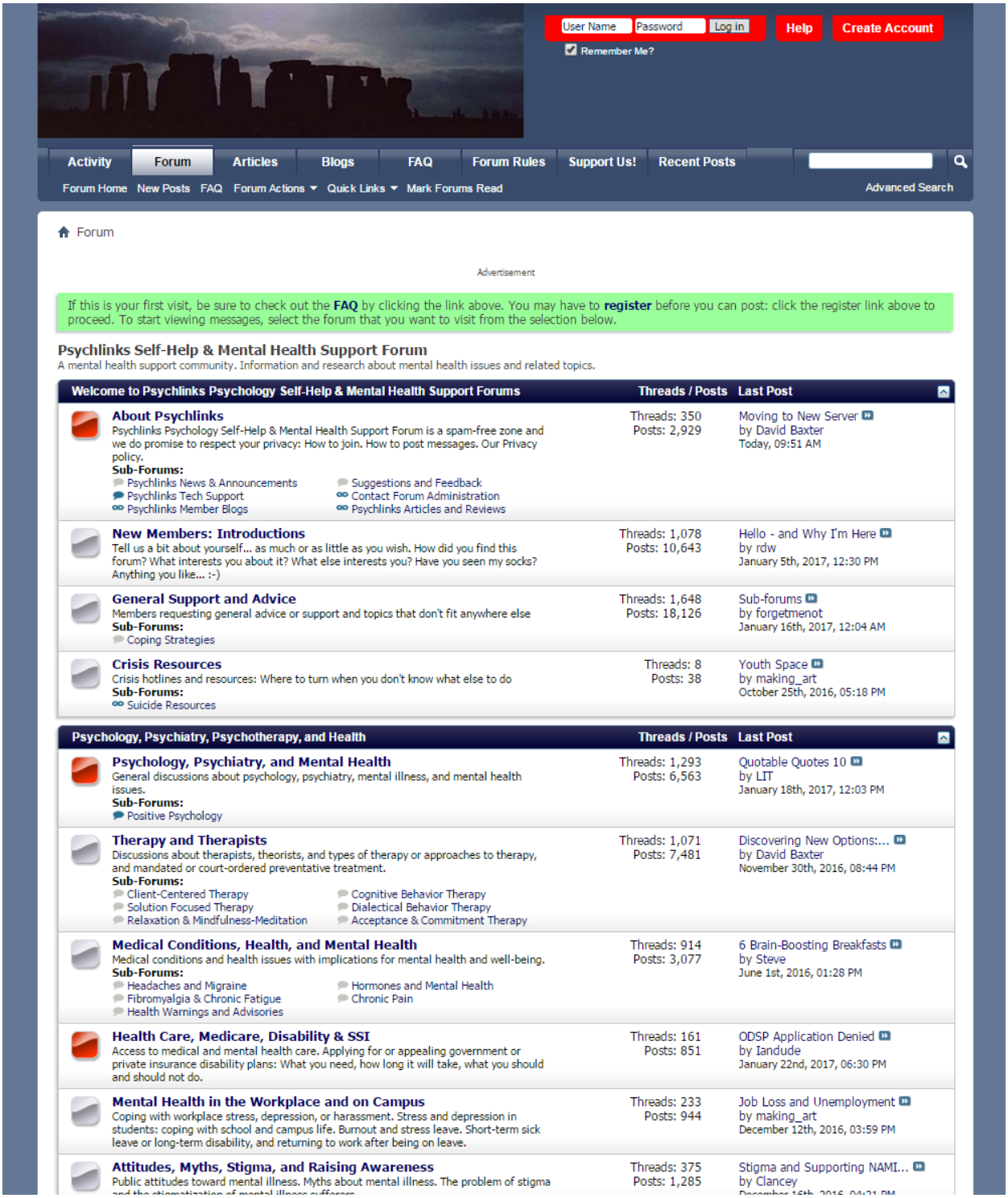


Figure 3.4: Screenshot of the front page of the PSY forum, retrieved on 25/01/17

threads, and then fetching posts from each thread. This procedure does not yield a perfect snapshot of the forums, as some posts contributed after the start of the crawl might have been included; however, such inaccuracies are extremely minor, since the crawl of even the largest forum required only a few days and for all four forums the number of posts per day is extremely small compared to the total post count (see Table 3.1). I did not retrieve any resource (e.g., I did not collect user avatars, or pictures embedded in posts) other than the raw HTML pages, which I did not store. Instead, I parsed all information besides post content, extracted metadata, and stored all information useful for the analyses into a PostgreSQL database. The schema for the database, which includes details on the fields for each forum entity I stored, is depicted in Figure 3.5. I parsed post content at a later time, for a better trade off between reduction of noise and loss of information in the data. The next section details this process of cleaning raw data.

3.5 Data curation

Raw data showed missing values, corrupted encoding, invalid content – as with all real-world data, they needed some curation before being useful. I discarded all dates that preceded the creation of the forums from posts and user profiles. I re-encoded all text to utf-8, attemping a cast to ascii for characters outside of the encoding. The post contents often were invalid HTML, or contained broken bbcode[12]. In fact, the post editor in most forums allows users to add rich text formatting, embed smileys, links, images and videos, and other forum-specific features. I chose to strip all complexity from post content and to keep only the text, after extracting quotes. In particular, I discarded nested quotes (quotes embedded within other quotes) from the HTML parse tree, to streamline analysis and simplify result interpretation. I then extracted from each quote the author of the quoted post (when specified), and the link to the quoted post (when specified). I then stored posts' and quotes' text as the concatenation of the string elements in the remaining HTML.

[12] https://en.wikipedia.org/wiki/BBCode

I tentatively re-linked quotes missing a link to the quoted post, based on the quote's text: more specifically, any such "orphan" quote was linked to the latest post preceding it (in the same thread – inter-thread quotes are extremely rare) whose text was a super-string of the quote's text, and whose author matched the user cited in the quote (when specified). This "text-based linking" was crucial because quote format changed over time in all four forums: while initially quotes only included the plain text of the quoted posts, the forums added relatively early in their history the option of referencing the quoted post's author – and only some time later that of explicitly linking the quoted post. This change was most likely the result of updates to newer versions of the forum front-end software.

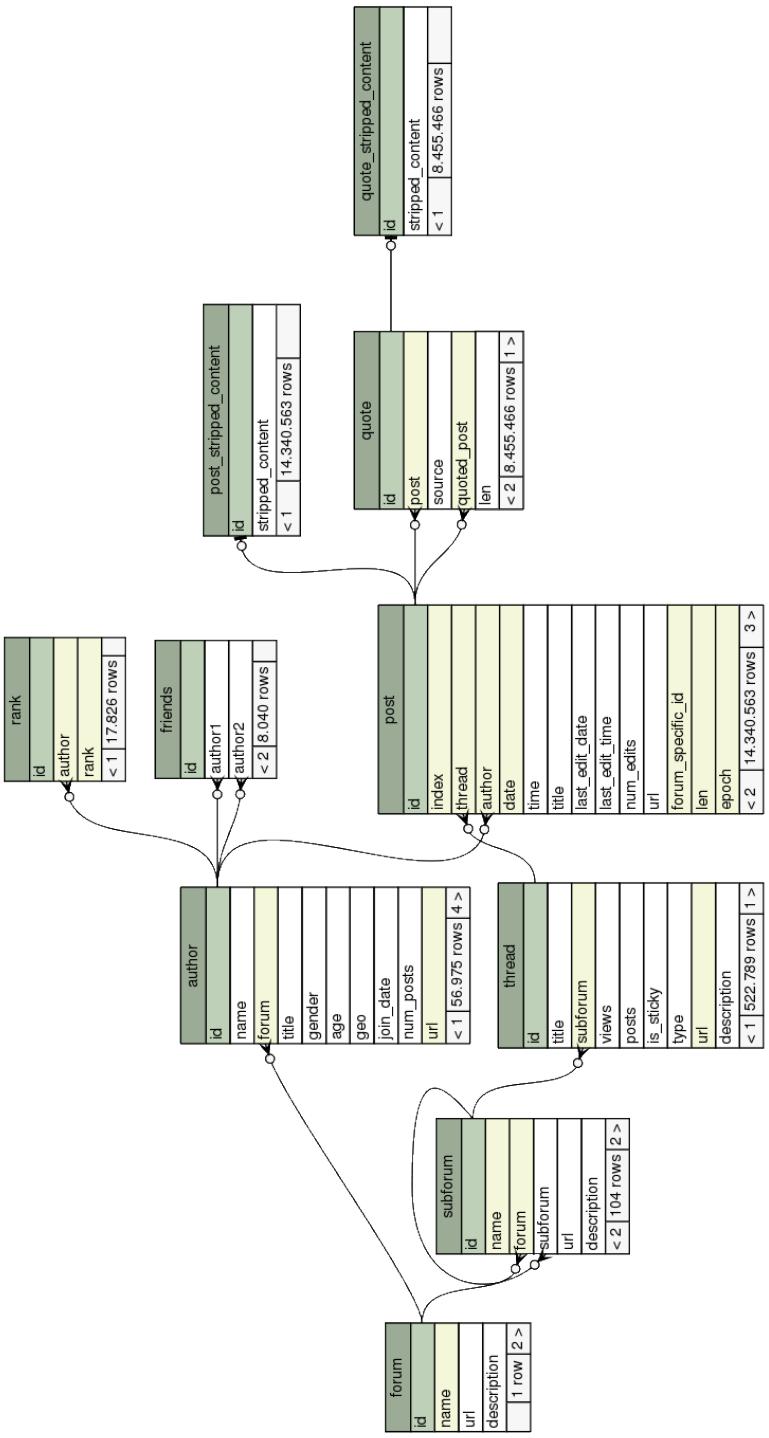


Figure 3.5: Database schema for the four forums. Each table corresponds to an entity in the forum (e.g. subforum, thread, user, etc.), and reports the metadata fields collected from the crawler.

3.6 Forum data in numbers

The following table shows some basic statistics of the forum data:

	RPG	**SWZ**	**TM**	**PSY**
posts	14.3M	1M	3.6M	0.15M
users	56.9K	29.9K	14.9K	2.8K
threads	522.7K	112.1K	49.2K	24K
quotes	8.4M	218.8K	1.6M	31.1K
timespan	'00-'13	'02-'14	'01-'14	'04-'14

Table 3.1: Overall data quantity for the four forums.

Data for all four forums spans a decade or more. The amount of data in terms of number of posts or users, however, varies by up to two orders of magnitude across forums. Also, it appears that users in different forums sport different levels of activity: the ratio of posts per user, threads per user, and posts per thread are all very different across forums. The size and diversity of the forums suggest that any coherent findings coherent across the four datasets are unlikely to be the result of overfitting.

3.7 Limitations

All analyses in this dissertation build upon data as processed in this chapter. However, data may be valuable beyond these analyses. Therefore, I would like to be very explicit about the limitations that come with the data gathering and curation process I followed.

I did not store the raw HTML files, and I committed to a predetermined choice of the metadata to store. This does not affect the analyses in this work; however, some information was never part of the dataset. Post footers and user signatures are missing, as well as avatars, and other multimedia resources in post contents. Also, I did not crawl the thread reply structure, as posts did not contain such information, and users could not see it by default.

Other information is stored as part of the raw post content but is removed during content curation. This information includes most notably all layout, typesetting, and formatting of post contents, as well as quote position within the quoting post, and nested quotes.

3.8 Privacy and ethical concerns

An additional cautionary word is due regarding the use of this dataset. All data gathered is public, with the possible exception of users' friend lists, which are publicly accessible after registration. However, forum data contain potentially sensitive information: users may disclose personal information within the context of the community they may not feel comfortable to reveal in other contexts. Data obtained from PSY, where users often seek help for a mental health condition, are an obvious example. This raises ethical concerns. All analyses in this dissertation are performed on

aggregate whenever possible, to preserve individual privacy. When that is not possible, for example when analyzing the role of individual users, all personally identifiable information is removed. Since there was no interaction with the users of the forums, this study did not require obtaining informed consent or board approval.

This dissertation bases its analyses on forum data. Forums may not be a "hip" venue for social computing research; however, their simple interface and rich data are well-suited to analyzing interaction patterns in online discussion. To account for the risks of overfitting, and biases coming from factors of scale, user base composition, and topic, this work concentrates on four, appropriately chosen, macroscopically different communities. The next chapter starts the analytical part of this dissertation, investigating the fundamental question: if one doesn't look at the *content* of our online interactions, does the *way* we interact reveal anything about us?

4
Identifying users through interaction

When two people talk face-to-face, they can learn about each other without saying a word. The way people communicate produces a constant stream of signals about them – this is a primal construct for organizing and coordinating socially. But what happens when people move their discussions to online social media? Does the way they communicate online still tell reveal something about them?

To my surprise, I could not find an existing framework to study interaction patterns in online discussion independently from content – in fact I could not even find an actionable definition for this concept. This chapter presents a study that addresses this gap[1]. It gives a general definition of interaction patterns as content-agnostic features of discussion. Then, it builds upon this definition to prove the foundations of the dissertation: content-agnostic features are personal signatures of user's interaction patterns in online discussion, and these signatures show consistent characteristics across different communities.

[1] This is joint work with Enoch Peserico, and was first presented in Samory et al., "Content attribution ignoring content", 2016

The customary approach to identifying users based on their style is through authorship analysis. Authorship analysis adapted text analysis techniques from before the digital era (e.g. stylometry) to the online context, to accurately identify users based on the content of their messages – Section 2.2 gives a more in-depth analysis of related results in the field. Similarly to authorhsip analysis, this work operationalizes identifying users through their style as authorship attribution and verification problems.

This work has the potential to overcome the limits of authorship analysis in online discussion. The typical text constructions used online have become shorter[2], possibly in an attempt to make the result more "engaging"[3]. Also, online text often deviates from literary language: vocabulary and grammar rapidly mutate and are replete with neologisms (e.g. "hashtags") and unconventional use of language (e.g. hashtags), often evolving into platform-dependent idioms (e.g. *chanspeak*). Interaction patterns may compensate, at least partially, for the reduction in quantity and "quality" of text.

[2] Alis et al., "Spatio-temporal variation of conversational utterances on Twitter.", 2013

[3] Facebook suggests to use *"short, fun-to-read copy and eye-catching images to get attention"*: `https://www.facebook.com/business/learn/facebook-page-create-posts`, accessed on 26/1/17

Moreover, this work may allow analyzing discussion where text is not present at all, which is an increasingly important venue of research. Sharing audio clips through instant messaging has become common practice. The primary content of several top-

traffic-driving platforms such as Pinterest and YouTube is visual. Social buttons such as Facebook's *"like"* and Google+'s *"+1"* are widely adopted as non-verbal manifestations of endorsement.

At a high level, the study bases on the assumption that all manually initiated interaction carries traits of its author, regardless of what has been shared. For practical applications, the hope is that these traits are present in online discussion even if the user interface does not make their presence or meaning explicit. This work shows supporting evidence through three contributions:

- *it provides an operative definition of* content-agnostic *features of communication;*
- *it proposes a case-study set of content-agnostic features for forum messages, and prove its effectiveness in dealing with authorship analysis tasks;*
- *it provides a preliminary taxonomy of content-agnostic features, analyzing and comparing the information content of different families of features.*

The rest of this chapter is organized as follows. In Section 4.2 I introduce the definition of *content-agnostic features* – in a nutshell, those features that can be measured even when each symbol in a contribution (e.g. every character in a text) is replaced with a copy of a standard symbol (e.g. a blank). I gather insights from discourse and social network analysis to extract 49 content-agnostic features regarding quantitative, temporal and relational traits of a post and its thread. I then provide an experimental assessment of the authorship information captured by content-agnostic features using them (and them alone) for two classification tasks: deciding if a given post in a discussion forum has been authored by a given user, and attributing a post to an author from a set of candidates. The experimental results on the four forums introduced in Chapter 3 are described in Sections 4.7 and 4.8: the first task can be performed with 77% accuracy; the second with 94% accuracy when attributing authorship to an author in a given pair. In Section 4.9 I analyze how individual features, and groups of features, affect classification performance, before concluding in Section 4.10 with a summary of the results, an analysis of their significance, and some possible directions of future work.

4.1 Research question

Before proceeding, I clarify that this work addresses two open questions:

RQ1 Do users leave recognizable traces in online discussion forums that do not depend on post content?

RQ2 If it is possible to predict user identity from content-agnostic features, does the prediction power of different features remain stable across different forums?

4.2 *Definition of content-agnostic*

Literature lacks a definition for features of a discussion that do not depend on content. This section defines these features *"content-agnostic"*.

Informally, a feature in a given online discourse is content-agnostic if it depends solely on how that content is produced, interlinked etc. Examples of content-agnostic features would be the levels of "burstiness" in post activity of a given thread in an online forum, or the topological properties of a reply/repost graph in a social network.

This concept can be made more formal by modeling an online discourse as a graph of elementary symbols (characters for textual discourses, pixels or images for visual ones etc.), where content fruition follows the graph's arcs: e.g. a hyperlinked text would be represented by long chains of symbols, with the occasional arc (a hyperlink) connecting different chains. Portions of the graph can be, and typically are, annotated with additional metadata (e.g. times of posting, "likes" etc.). A feature of the discourse, or of an individual portion of the discourse, is content-agnostic if it can still be computed from a modified version of the annotated discourse graph in which every symbol has been replaced by a copy of a standard "blank" symbol.

Note that this definition is not completely rigorous (something which would require a much more complex modeling of discourse), but it still yields a simple, and in most cases objective, criterion to assess whether a feature is content-agnostic. This is more evident as soon as the discourse abstraction is given a particular, concrete form – e.g. an online discussion forum, for which content-agnostic features are those that can be computed replacing every character in the forum's posts with a blank (Figure 4.1).

Figure 4.1: Visualization of content-agnostic features in a hypothetical Facebook wall post. A feature is content-agnostic if it can be measured after extracting metadata, and replacing all text with an "X" and all image pixels with a predefined color.

I remark that, although content-agnostic features correspond somewhat loosely to the informal notion of "structural" features in

authorship studies (see Section 2.2), they differ in two important respects. First, they do not depend on any property of the writing style such as period length or capitalization of words. Second, they deal with a more general class of online discourses, involving links, multimedia content, social buttons etc.

The following section makes the notion of content-agnostic features more concrete, by providing an example set of such features in the context of online forums. The information content of this feature set is then assessed in Sections 4.7, 4.8, and 4.9.

4.3 Content-agnostic features for forum posts

In the context of online discussion forums I identify a set of 49 content-agnostic features that may guide authorship analysis of an individual post. I organize these features in a simple taxonomy to more easily analyze their role. This taxonomy has two axes: *scope* and *type*.

The scope of a feature can be *post*, for features that look only at the immediate surroundings of a post (e.g. the post itself, the posts immediately preceding it, and cited posts); and *thread*, for features that characterize the post's discussion thread in its entirety (thread feature values are shared by all posts within the same thread). This distinction allows verifying to what extent information "local" to the contribution is sufficient to identify the author fingerprint.

The type of a feature can be *intensity*, *time*, or *link*. Intensity features quantify posting volume and curation effort. Time features assess timing, both in absolute terms and relative to other posts and threads. Link features measure various aspects with a "social" valence (e.g acknowledgement, attribution, and endorsement of other posts).

4.4 Taxonomy of content-agnostic features

Scope and type for each feature are listed in parentheses immediately after the feature's description. Features that are different aggregates on the same metric (such as `f36-38`) are grouped together. The distribution of features per scope and type is presented in Table 4.1.

`f1`: time of posting, in minutes since Jan 1, 1970 *(post, time)*

`f2`: if the post has quotes *(post, link)*

`f3`: number of quotes in the post *(post, link)*

`f4`: number of distinct posts quoted in the post *(post, link)*

`f5`: number of distinct authors quoted in the post *(post, link)*

`f6`: if the post quotes a single other post multiple times *(post, link)*

`f7:` average fraction of quoted posts' characters that are quoted *(post, link)*

`f8:` if the post's title contains a tag *(post, intensity)*

`f9:` if the post is the first in the thread *(post, intensity)*

`f10:` day of week of posting *(post, time)*

`f11:` time of day of posting, in minutes *(post, time)*

`f12:` month of posting *(post, time)*

`f13:` day of year of posting *(post, time)*

`f14:` day of month of posting *(post, time)*

`f15:` if the post has been edited *(post, intensity)*

`f16:` if the post contains links *(post, intensity)*

`f17:` if the post links to resources external to the site *(post, intensity)*

`f18:` length ratio between the post and its quotes *(post, link)*

`f19:` post's number of characters, excluding quotes *(post, intensity)*

`f20:` cumulative number of characters of the post's quotes *(post, link)*

`f21:` time difference since the previous post in the thread sequence, in minutes *(post, time)*

`f22-24:` average, maximum, minimum time difference between consecutive posts in the thread, in minutes *(thread, time)*

`f25-27:` average, maximum, minimum time difference between consecutive posts by different authors in the thread, in minutes *(thread, time)*

`f28-30:` average, maximum, minimum time difference between a quote to an author, and the next post by that author in the thread, in minutes *(thread, time)*

`f31:` fraction of posts in the thread that quote an author and are immediately followed in the thread being by that author *(thread, intensity)*

`f32-34:` average, maximum, minimum number of authors between two consecutive posts in the thread by any author *(thread, intensity)*

`f35:` number of different authors quoted by all posts in the thread *(thread, link)*

`f36-38`: average, maximum, minimum number of posts in between two consecutive posts in the thread by any author *(thread, intensity)*

`f39`: total running time of the thread, in days *(thread, time)*

`f40`: number of different authors in the thread *(thread, intensity)*

`f41`: number of posts in the thread *(thread, intensity)*

`f42`: average number of characters of post in the thread *(thread, intensity)*

`f43`: index of the first posts by each author in the thread, averaged *(thread, intensity)*

`f44`: post index (sequential number in the thread) *(post, intensity)*

`f45`: average time difference since thread start of the first posts by each author in the thread, in minutes *(thread, time)*

`f46`: time difference since thread start, in minutes *(post, time)*

`f47`: average time difference between the last 10 posts in the thread sequence, in minutes *(post, time)*

`f48`: thread number of views *(thread, intensity)*

`f49`: ratio between the number of posts and the number of views of the thread *(thread, intensity)*

It is important to clarify that this case-study feature set is far from being exhaustive – it could be extended, for example, incorporating "social" attributes, such as the popularity of the author, or frequent commentators to the author's posts (as suggested by an anonymous reviewer of this work).

Moreover, this feature set is tailored to represent interaction within an online forum. However, I stress that the definition of content-agnostic features applies to a much wider spectrum of online interactions. In fact, this taxonomy is based on principles that generalize easily, and it should simplify expanding the feature set and/or porting it to different contexts.

	post	*thread*	*total*
intensity	7	13	20
link	8	1	9
time	9	11	20
total	24	25	49

Table 4.1: Number of features per scope and type.

4.5 Method

This work approaches authorship analysis as a supervised learning problem. It uses a standard classification setup (see Figure 4.2), consisting of feature extraction, data sampling, cross-validated model training and evaluation. This section discusses each of these processing steps in detail.

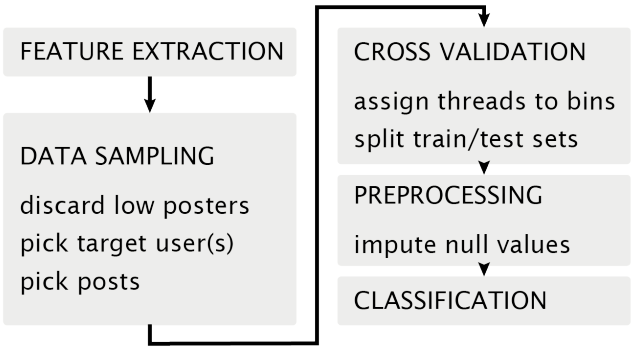


Figure 4.2: Experimental setup for authorship verification and attribution using content-agnostic features.

4.5.1 *Sampling*

I limit target authors to users with more than 100 posts who contributed to more than 10 threads, to reduce noise. Table 4.2 shows that this filtering preserves the vast majority ($85\% - 97\%$) of forum posts, even though it does eliminate a large number of "occasional" authors.

	RPG	**SWZ**	**TM**	**PSY**
posts	14.3M	1M	3.6M	0.15M
users	56.9K	29.9K	14.9K	2.8K
$posts_f$	13.9M	3.5M	0.89M	0.13M
$users_f$	7.8K	2.1K	847	125

Table 4.2: Number of posts and users retained after filtering out users with few posts (filtered quantities have a subscript f).

I fully acknowledge that the filtering threshold is somewhat arbitrary: determining the minimum number of contributions per author below which author identity is effectively drowned by noise is an interesting open problem – the answer arguably depends on the feature set[4] and on the number of authors.

[4] Eder, "Does size matter? Authorship attribution, small samples, big problem", 2014

4.5.2 *Learning pipeline*

Feature extraction is straightforward, as links between threads, posts, authors, and quotes are parsed during crawling (detailed in Section 3.4), and stored in a database. Section 4.3 provides the complete list of features used. I substitute missing feature values with a fixed out-of-range placeholder, to meet common prerequisites for a variety of classifiers. Preliminary tests suggest that elaborate preprocessing yields very marginal accuracy gains, at the price of substantial additional complexity.

I use a Random Forest classifier[5] to learn author profiles[6]. Random Forests are an ensemble method that outputs the mode of the classes predicted by a number of decision trees. The trees are trained on distinct random samples with replacement of data (*bootstrap samples*). At each step in the learning process, a tree considers a random sample of the features to find the best node split. This procedure decorrelates decision trees in the forest, thus reducing

[5] `http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`

[6] Breiman, "Random forests", 2001

the variance of the model. Each training sample P_i is then used to compute an estimate of the generalization error (*out-of-bag error*), averaging the prediction error on trees that did not have P_i in their bootstrap sample.

One crucial advantage of Random Forests is that they can provide intuitive measures of feature importance. One such measure is *mean decrease impurity*, that assesses the total decrease in node impurity due to splits on a given feature, weighed by the proportion of samples routed to each node. Another is *mean decrease accuracy*, that assesses the normalized misclassification rate when the values for each feature are randomly permuted.

Recent work successfully applied Random Forests to authorship analysis problems[7]. I have chosen Random Forests over other widely used classifiers because of their inherent feature evaluation capability, their minimal tuning requirements, and their out-of-the-box performance. We employ the implementation provided by `scikit-learn`[8], using 200 estimators and information gain as a splitting criterion. I note in passing that I have cross-checked our findings on alternative models such as Decision Trees and Support Vector Machines: results (not presented in this work) are of lower accuracy, albeit qualitatively comparable.

[7] Abdallah et al., "Detecting Email Forgery using Random Forests and Naïve Bayes Classifiers", 2012; Arakawa et al., "Adding twitter-specific features to stylistic features for classifying tweets by user type and number of retweets", 2014; Pratanwanich et al., "Who Wrote This? Textual Modeling with Authorship Attribution in Big Data", 2014

[8] Pedregosa et al., "Scikit-learn: Machine Learning in Python", 2012

4.5.3 *Performance metrics*

I measure prediction performance using k-fold cross-validation. However, splitting data into each fold requires some care, since posts in a given thread all share *thread*-level feature values (Section 4.3). Thus, I do not randomly assign posts to folds, as this could make information from the validation set available during training. Instead, for each target author, I pre-emptively assign a random k-partition of threads to the k folds, and then pick posts from each thread. I set $k = 10$, equal to the minimum number of threads per user, so that all target authors have at least one post available for each fold. Note that if one considers all posts by a given author, the number of posts may vary from fold to fold. An alternative setup would be to train two models, one on *post*-level and one on *thread*-level features - however, this approach is cumbersome and would seriously limit the ability to explore relationships between different features (see Section 4.9).

I assess the performance of the classifier on each cross-validation round. Then, I compute the average metrics per author, and average the result over all the authors of each (forum) dataset. The final metrics are therefore macro-averages on the cross-validation rounds. As sample size in each cross-validation round may vary, I also compute the *global* accuracy for each dataset, gathering all predictions, and evaluating the total fraction of correct classifications.

4.6 Textual baselines

I benchmark content-agnostic features against three content-dependent feature sets, and against a combination of the three. These simple yet widespread feature sets are at the core of most stylometric approaches to online authorship analysis. Also, I evaluate the performance of content-agnostic and content-dependent features combined. This puts into perspective how much information is captured by content-agnostic features alone. I intentionally do not perform sophisticated "data massaging" or model tuning, as the goal of this work is to prove the general applicability of content-agnostic features to authorship analysis (i.e. if content-agnostic features carry significant information of user interaction patterns), rather than sheer classification accuracy.

The three content-dependent feature sets are: character *trigrams* (the frequency in a post of the most common sequences of three characters in all posts), word unigrams (the frequency in a post of the most common words in all posts – from now on *bag-of-words*), and term frequency-inverse document frequency (a bag-of-words that penalizes words that are frequent in all posts, and thus less informative [9] – from now on *tf-idf*). I use 100-dimensional vectors for each feature type. I apply minimal text preprocessing: I substitute all non-letter characters with whitespace, convert text to lower case, and eliminate stop words (using the `nltk` package[10]). These features, and their combinations, are evaluated on the same data and on the same train/test splits as the content-agnostic features.

[9] Manning et al., *Introduction to Information Retrieval*, 2008

[10] `http://www.nltk.org/`

4.7 Authorship verification

I frame the task of identifying a single user's set of contributions as a classification problem, formalized as:

Authorship verification: *Given access to all posts in the training set, and given a post p from the validation set drawn uniformly at random with probability $\frac{1}{2}$ from those authored by $\mathcal{A}$, and with probability $\frac{1}{2}$ from those not authored by $\mathcal{A}$, determine if p was authored by $\mathcal{A}$.*

Note that this formulation corresponds to a binary, balanced classification problem. Enforcing a probability equal to $1/2$ that the post's author is $\mathcal{A}$ (rather than a probability proportional to the fraction of posts of $\mathcal{A}$ in the corpus) allows for easier interpretation compared to a baseline "coin-flipping" strategy (that outputs "$\mathcal{A}$" or "not $\mathcal{A}$" each with probability $1/2$ without looking at the post). As suggested by previous literature (e.g.[11]), we allow the classifier to train both on posts authored by $\mathcal{A}$ and on posts not authored by $\mathcal{A}$.

I tested classification performance for 100 randomly sampled users per forum, using all posts by the target author, and sampling for each fold an equal number of posts by an "impostor" that is effectively the collective of all other users (including users with few posts).

[11] Brocardo et al., "Authorship verification of e-mail and tweet messages applied for continuous authentication", 2014; Koppel et al., "The "Fundamental Problem" of Authorship Attribution", 2012; Koppel et al., "Measuring differentiability: unmasking pseudonymous authors", 2007

	RPG	TM	SWZ	PSY
accuracy	0.79 0.11	0.72 0.15	0.75 0.13	0.77 0.11
precision	0.83 0.15	0.79 0.25	0.79 0.20	0.81 0.15
recall	0.71 0.22	0.54 0.31	0.64 0.25	0.70 0.21
F1	0.75 0.19	0.60 0.29	0.69 0.22	0.73 0.18
AUC	0.89 0.09	0.85 0.15	0.85 0.12	0.86 0.11
global accuracy	0.75	0.65	0.78	0.76

Table 4.3: Average and standard deviation (in gray font to the right) for various metrics of author verification.

The average and standard deviation of various classification metrics are presented in Table 4.3. Classification accuracy, averaged across all datasets, is 76%. Other classification metrics yield similar results. Precision is greater than recall in all cases: while the classifier predicts the author class less frequently, when predicted it is more likely to be correct. When a metric is inapplicable (e.g., precision when a class is not predicted), we set its value to 0, so as to present a "conservative" performance analysis. Using all features to split nodes, while capping growth of trees in the Random Forest, yields both faster training times and a greater balance between precision and recall.

I investigated the variability of accuracy values for all authors under consideration (standard deviations are roughly $10 - 12\%$). *Global* accuracy, i.e. accuracy averaged over all posts, is often lower than the *macro-average*, over all users, of accuracy averaged over posts by that user. This suggests that users with many posts might be more difficult to classify. Indeed, for authors with more than 500 posts, post count exhibits a mild negative correlation with accuracy for all datasets (Pearson's $r \in [-.08, -.32]$). This could be due to extremely prolific authors exhibiting a variety of interaction styles – indeed several of these authors are "virtual" users that do not correspond to a single person (such as "Rpg.net's Reviews"). Another hypothesis is that users with a long contribution history may change interaction patterns over time, fuzzying their classification profile.

To test the second hypothesis, I performed a simple experiment. I sampled 50 users with more than 500 posts from each dataset. We then divided author posts into three sets: $P^{(0)}$ (all of an author's posts in his first three months on the forum) $P^{(1)}$ (all of an author's posts in the next three months), and $P^{(2)}$ (all of an author's posts in the three months starting one year after his first post). In a Wilcoxon's signed-rank test, $\alpha = 0.05$, I found that training the classifier on $P^{(0)}$ results on average on lower prediction accuracy for $P^{(2)}$ than for $P^{(1)}$. This suggests that a user's "interaction profile" does indeed evolve over time.

I now compare content-agnostic and content-dependent features. Content-agnostic features consistently outperform in all classification metrics the content-dependent baseline presented in Section 4.6 – in increasing order of performance, 100-dimensional trigrams, bag-of-words, tf-idf features, and their combination. Figure 4.3 shows classification metrics for the RPG dataset; results for the

other three datasets are similar.

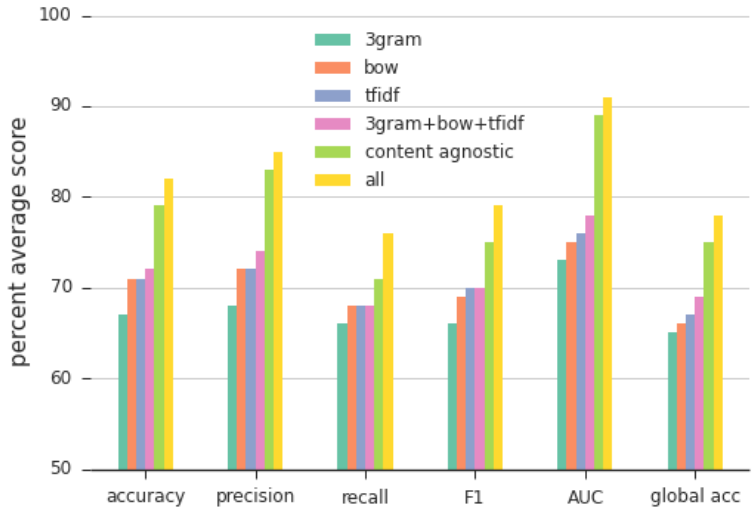


Figure 4.3: Average classification metrics for the authorship verification task, on the RPG dataset, considering content-agnostic features, content-dependent features (trigrams, bag-of-words, tf-idf, and their combination), and the combination of all features.

Unsurprisingly, the most effective approach (albeit by a small margin) is to combine content-agnostic and content-dependent features. This shows that content-agnostic features are not only a feasible alternative to content-dependent features (e.g. when the latter are difficult or impossible to extract), but also an effective complement to them to boost classification accuracy.

4.8 *Authorship attribution*

One can formalize the task of deciding which of n posters is the author of a given post as:

Authorship attribution: *Consider an author set of n authors $\mathcal{A}_1, \ldots, \mathcal{A}_n$. Given access to all posts in the training set, and given a post p from the validation set drawn uniformly at random with probability $\frac{1}{n}$ from those authored by $\mathcal{A}_i$ (for $1 \leq i \leq n$), determine which of $\mathcal{A}_1, \ldots, \mathcal{A}_n$ is the author of p.*

This formulation is that of an n-class, single label, balanced classification problem. As in the case of author verification, we enforce an equal probability of drawing a post by any given author within the $n-$author set, regardless of the total number of posts by that author in the corpus. This makes results more easily interpreted; in particular, it allows immediate comparison to a baseline classifier that attributes a post to an author chosen uniformly at random in the author set (thus producing a correct attribution with probability $1/n$). I apply the same basic setting explained in Section 4.5, to test how classification performance varies increasing the number of authors[12]. The difference from the authorship verification setup in Section 4.7 is that I randomly sample n authors, $n \in \{2, 5, 10, 20, 50\}$. I sample posts as follows: for each author, I

[12] Zheng et al., "A framework for authorship identification of online messages: Writing-style features and classification techniques", 2006

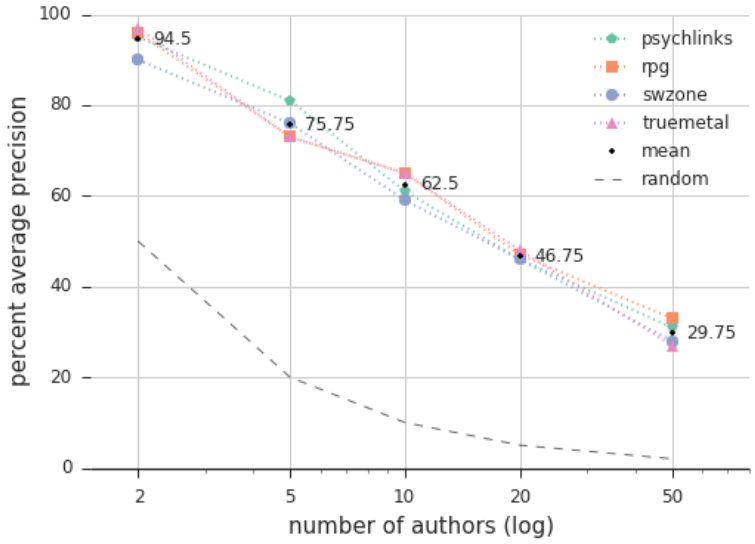


Figure 4.4: Average precision versus number of authors for the authorship attribution task, for all datasets. The average values are reported in text, next to the individual results. The lowest curve shows scores for the random attribution baseline.

partition his threads into folds; then, for each fold i, I compute the minimum number of posts per author p_i, and I assign to the fold p_i random posts from each author. For each value of n, I repeat the experiment 10 times, to stabilize results. I measure classification accuracy as *average precision*, i.e. the overall fraction of correctly attributed posts.

Accuracy is remarkably high, above 94% for 2 authors, and 75% for 5 authors averaged across all datasets, consistently beating the random baseline by a large margin in all cases. The global accuracy values never depart from the macro-averaged ones by more than 2%, and are therefore omitted. Accuracy variation between different forums is also minor, validating the hypothesis of robustness to language, topic, and community size.

Error increases with the number of authors, albeit slowly – apparently logarithmically (see Figure 4.4). This degradation of performance is, on the one hand, natural (with more authors, the average "distance" between them in feature-space becomes smaller and errors more likely), and has been reported by previous work on authorship attribution (see e.g. Juola[13]). On the other hand, specialized, qualitatively different approaches may be used to address large-scale authorship analysis[14].

Content-agnostic features compare favorably to this simple content-dependent baseline, surpassing the individual content-dependent feature sets, and their combination, by fairly large margins – as in the authorship verification task. Figure 4.5 shows average precision scores versus number of authors for the RPG dataset; the relative performance for all other datasets is similar. Content-agnostic and content-dependent features exhibit less synergy for authorship attribution than for authorship verification, and using only the former produces no appreciable loss of precision com-

[13] Juola, "Authorship Attribution", 2007

[14] Koppel et al., "Computational Methods in Authorship Attribution", 2008; Narayanan et al., "On the Feasibility of Internet-Scale Author Identification", 2012

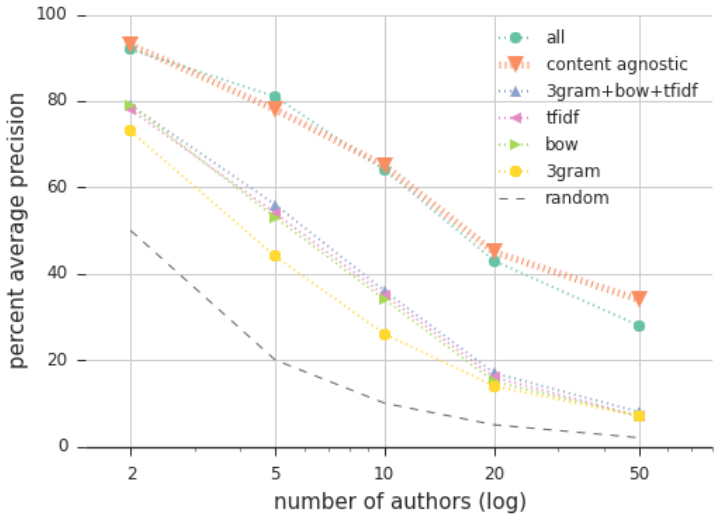


Figure 4.5: Average attribution precision versus number of authors for the RPG dataset, considering content-agnostic, content-dependent (trigrams, bag-of-words, tf-idf, and their combination), and the combination of all features, as well as the random baseline.

pared to using both – in fact, sometimes less than then effects of noise in training and cross-validation. The accuracy values reported are in line with previous literature on author attribution in online forums: for 5 authors Mohtasseb et al.[15] report 70 – 98% accuracy depending on post length, Pillay et al.[16] 69 – 91% depending on the classifier and the feature set used, and Abbasi et al.[17] 72 – 97% depending on the classifier used and on post language. Note however that *this is the first work to perform authorship attribution without looking at post content.*

[15] Mohtasseb et al., "More blogging features for author identification", 2009

[16] Pillay et al., "Authorship attribution of web forum posts", 2010

[17] Abbasi et al., "Applying authorship analysis to extremist-group web forum messages", 2005

4.9 *Feature performance*

I showed that content-agnostic features (section 4.3) collectively perform well in authorship analysis of forum posts (sections 4.7 and 4.8). In this section I discuss which features are the most important in identifying a user, considering the average feature weight assigned by the classifier in solving authorship verification problems (Section 4.9.1). I then broaden these considerations to feature *groups* combining the weights of the individual features (Section 4.9.2).

Note that feature correlations may influence weights: given two highly correlated features, the classifier might assign a high weight to one, and almost no weight to the other, or viceversa. However, this effect should be mitigated by the Random Forest's random feature subset choice at each node, and by the iterated model training for each author.

Feature weights reported in this section are computed using the *mean decrease impurity* criterion (see Section 4.5). Computing feature weights using *mean decrease accuracy* on the test sets leads to similar, qualitatively equivalent results: while the distribution is flatter, top-ranking features match, and group-wise relationships are maintained.

4.9.1 *Performance of individual features*

feature	*weight*	f29	0.0205
f1	0.1042	f21	0.0204
f49	0.0739	f33	0.0203
f11	0.0479	f44	0.0195
f13	0.0365	f14	0.0161
f42	0.0335	f18	0.0160
f41	0.0301	f30	0.0152
f22	0.0282	f19	0.0144
f48	0.0279	f20	0.0136
f25	0.0267	f7	0.0110
f32	0.0263	f10	0.0109
f40	0.0257	f27	0.0109
f36	0.0242	f8	0.0090
f43	0.0239	f24	0.0089
f23	0.0232	f16	0.0075
f12	0.0231	f4	0.0051
f45	0.0227	f3	0.0048
f26	0.0226	f17	0.0044
f37	0.0226	f38	0.0043
f46	0.0226	f2	0.0039
f35	0.0215	f6	0.0034
f31	0.0211	f5	0.0033
f39	0.0210	f34	0.0032
f28	0.0206	f15	0.0017
f47	0.0206	f9	0.0008

Table 4.4: Average feature weights for the RPG dataset

Feature ranking according to weight is consistent across the four forums, with pairwise Kendall's $\tau > 0.5$, $p < 10^{-5}$, and group-wise Kendall's $W > 0.8$. Since features are robust and maintain their role across the four datasets, we focus on the case of the *RPG* dataset (Table 4.4).

The top 10 features account for 44% of the cumulative weight, and are highly varied in both category and scope.

`f49`, the ratio between the number of posts and views of the thread, and `f42`, the average post length in characters in the thread, suggest that users choose threads according to how well the discussion motivates viewers into being active participants, and to how much effort participants devote to their posts.

`f1`, `f11` and `f13`, the absolute time, time of day, and day of year of posting, suggest that regular users effectively develop routines that make their interaction predictable.

Note that `f1`, the absolute time of posting, is the feature with the heaviest weight. Given the 10+ year timespan of our datasets, one might then wonder if the accuracy of our content-agnostic verification depends mostly on rejecting as "impostor" posts outside of the forum lifetime of the main author – e.g. rejecting a post from 2005 if the author's other messages are all posted after 2010. This

is not the case (note that after filtering out "occasional" posters, remaining ones all tend to have fairly long lifetimes). I repeated the experiments only selecting "impostor" posts from the timeframe the main author was active in, without observing any significant drop in accuracy. Even removing `f1` entirely from the feature set results in a very small drop in verification accuracy (from 76% to 74%). In fact, upon visual inspection, it seems that `f1` may exploit the bursty nature of user activity: users posting peaks at relatively regular intervals (e.g. one user posts several times in one hour, then waits one day before posting again), but the phase and period of those intervals is different between users (e.g. some users post weekly every Monday, while others every three days regardless of the day of the week). Exploring this issue further is certainly a promising direction of future research.

Contrary to expectations, the post sequential number in the thread (`f44`), and the thread opening post indicator (`f9`) have low weights. This might be due to correlation with other features, e.g. `f44` is highly correlated with `f41`, the number of posts in the thread.

In brief, it appears that what makes users identifiable against impostors are the routine, bursty nature of human communication, and the level of engagement provoked by the chosen discussions. The next section gives context to this analysis observing features aggregated into their respective category.

4.9.2 *Performance of feature taxa*

Section 4.2 provided a taxonomy for the features, categorizing them by type and scope, to obtain a clearer high-level picture of the main drivers of overall performance (Figure 4.6). The most influential feature types are *time* (encompassing features like the absolute time and the time of day of posting) and *intensity* (encompassing features like the number of posts compared to the number of visualizations of the thread). The aggregate weight of time and intensity features is respectively 0.52 and 0.39.

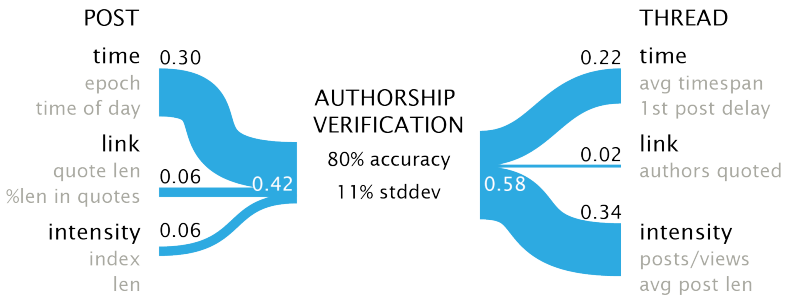


Figure 4.6: Cumulative feature weights by feature scope and type for the RPG dataset. Top features are reported below each category.

Interestingly, although all four datasets exhibit heavy use of quotes (see Table 3.1 and Chapter 5) and 9 out of 49 features are categorized as *link*, this feature type exhibits relatively low predic-

tive power. The highest-ranked feature in this category (`f35`, the total number of authors quoted in the thread) ranks 20^{th} overall. It is possible that an expanded sample of *link* features could improve their relevance.

Looking at features divided by scope, *thread*-level and *post*-level features have approximately the same aggregate weight. This means that context adds valuable information about a user's contribution patterns. An explanation for the performance of *thread*-level features might be that users have a selection bias for threads with specific characteristics. On the other hand, features "local" to the post retain a significant fraction of predictive power.

I also investigated feature importance in combinations of scope and type, repeating the classification experiment, using only features in each feature subset. The average accuracy results are reported in Table 4.5 - while each result comes from a different random sample of users, reiterations confirmed the findings. Surprisingly, using only *post*-level features leads to a classification almost as accurate as using all features. In particular, using only the intersection of *time*-based and *post*-level features reaches almost the same results: the accuracy averaged on all datasets is 76%. This means that simpler, more efficient classifiers could be trained on this restricted set of 9 features.

	post	*thread*	*total*
intensity	0.67 0.10	0.68 0.12	0.71 0.12
link	0.63 0.09	0.57 0.12	0.64 0.11
time	0.77 0.10	0.63 0.12	0.77 0.11
total	0.79 0.10	0.69 0.12	0.79 0.11

Table 4.5: Average accuracy for each content-agnostic feature group for the RPG dataset. The small text is the standard deviation on the measures.

4.10 Discussion

It is not just *what* we contribute that defines our online identities, but *how* we do it. A simple set of 49 post features *completely independent of post content* identifies authors of forum posts with accuracy comparable to standard stylometric approaches. Furthermore, accuracy appears remarkably stable across a spectrum of forums sporting widely different memberships, topics of discussion, and interaction patterns. It appears that what makes users identifiable are their routine, bursty activity[18], and the level of engagement provoked by the chosen discussions.

I am not claiming this approach to authorship analysis is better than content-based ones, or that content should simply be ditched in favour of content-agnostic features. However, this result show that there is a wealth of authorship information outside of actual content, in interaction patterns. Interaction patterns therefore could be used *in addition* to content-embedded information to improve authorship analysis – or could serve as a substitute when text content is not available.

[18] which has been found to be a fairly general property of human communication: Barabási ("The origin of bursts and heavy tails in human dynamics")

4.11 Implications

This section summarizes what is the impact of this work, and how it can inform future research and applications.

4.11.1 Theoretical implications

This work introduces an empirical, actionable definition of interaction patterns as content-agnostic features, and it proves they are *per se* signatures of how users interact in online discussion. This is a basic result, however it is a cornerstone onto which research may build a better understanding of user behavior and online discussion.

With respect to the individual features proposed in this study for forum post authorship, it seems that some features are consistently more predictive across all four communities. This suggests that these features do not depend on the type, language, size, or focus of the community – instead, they seem representative of the way users interact in general (at least, on forum-like discussion platforms). Interaction patterns may be a key element in studying online behavior beyond the limits of single platforms.

With respect to the proposed feature categories, it seems features on the entire discussion are informative of user identity. Note that the target user cannot directly manipulate these features, since they come from all participants to the discussion: it is likely that what thread-level really measure is the decision of the user to take part to the discussion or not. This suggests interaction patterns may also capture social signals, as already suggested by research on tie strength in online media.

4.11.2 Practical implications

It is possible to identify social media users without looking at the content they produce. A classifier trained only on a few tens of features that are local to a post can attain relatively high accuracy. On the one hand, it appears efficient authorship analysis tools can be built with minimal content disclosure requirements. on the other hand, this has serious implications for the way we model and perceive online privacy[19]. For instance, it is a common practice to share data anonymizing personally identifiable information in content. This approach would not protect user identity from being revealed through content-agnostic features. In fact, even end-to-end encryption recently deployed to instant messaging application could not protect from content-agnostic analyses[20].

[19] Montjoye et al., "Unique in the Crowd: The privacy bounds of human mobility.", 2013

[20] A recent article discusses why this may be an issue for all current major instant messaging applications, in the face of government surveillance: `https://medium.freecodecamp.com/e93346b3c1f0`, accessed on 26/1/17

4.11.3 Future work

Results in this work could certainly be improved and extended. For example, I only consider a small feature set; more extensive feature engineering may well improve classification accuracy. In particular,

a different approach to accounting for social structure (the rationale behind the inclusion of *link* features) could prove effective[21]. The next chapter of this dissertation shows further supporting evidence.

Content-agnostic analysis should, by its very definition, be an easily "portable" tool. In this sense, it would be interesting to apply it to structurally different platforms, particularly those where the prevalence of non-textual information has so far prevented or limited any authorship analysis – e.g. tumblr, Pinterest, and Instagram. In fact, it would be extremely interesting to evaluate the performance of content-agnostic features for authorship analysis *across* platforms: do users have *unique* fingerprints that are maintained when moving conversation e.g. from Facebook to Twitter, or from offline to online? In a joint work with Chandrasekharan et al., I investigate a possible learning framework for addressing this problem – although not directly applying it to interaction patterns[22].

Finally, this work showed that interaction patterns capture signals of user identity; however, what kind of signals? How are these signals different from those captured through message content? Can interaction patterns characterize users, in addition to identifying them? Chapter 6 explores this direction.

[21] Govindan et al., "Local Structural Features Threaten Privacy across Social Networks", 2013; Koessler Gosnell, "Social Fingerprinting : Identifying Users of Social Networks by their Data Footprint", 2014; Narayanan et al., "De-anonymizing social networks", 2009

[22] Chandrasekharan et al., "The Bag of Communities Approach : Identifying Abusive Behavior Online with Preexisting Internet Data", 2017

ONLINE DISCUSSION challenges traditional content analysis techniques. This work introduces a research frame for studying users in online discussion through the lens of interaction patterns, completely disregarding content. To this end, it proposes an actionable definition of content-agnostic features. Then, it proves that content-agnostic features are a signatures of how users interact in online discussion. An out-of-the-box model trained on 49 content-agnostic features confirms the author of a message with 77% accuracy, and distinguishes between two users with 94% accuracy in four forums – comparably to textual baselines. An inspection of the feature weights shows that their role remains consistent across the four forums, suggesting that interaction patterns generalize well beyond an individual social medium. The proposed features best identify users through temporal aspects of their posts, and through the relative engagement provoked by the discussions they participate in.

This chapter showed that interaction patterns reflect the users as individuals. Next, the following chapter extends this research beyond individual users. It focuses one interaction medium, the quote, and analyzes what it reveals about the relationship between users, the user roles in the discussion community, and about the discussion community as a whole.

5
Modeling discussion communities through interaction

The previous chapter demonstrated that interaction patterns are signatures of how users interact in online discussion. However, interactions do not happen in a vacuum. Interaction patterns come from a consistent interplay between users and context: be it other content, other users, or the community. It comes natural to question whether they also tell something about the context. How do interaction patterns relate to the way people use the platform for discussing? How to the relationship between users? How to the structure of the discussion community? Or do the different users' interaction signatures meaninglessly juxtapose when they share the same discussion medium?

At a very high level, the underlying question is if interaction patterns are shared conventions, or mere accidents of users' "motor skills" in online discussion. This chapter presents a study that addresses this question[1]. In particular it concentrates on one mode of interaction, the quote, and uses it as a metric to measure activity in the four forums presented in Chapter 5. Quotes are excerpts from previous posts that a new post can cite.

But why concentrate on quotes? First, many online platforms feature quotes in various forms. One can consider replies, mentions, retweets, shares, repins as lower resolution versions of quotes, which allow a more fine-grained interaction with the content they refer to[2]. Therefore, findings on quotes should easily generalize to platforms other than forums.

Moreover, quotes are a rich, multifaceted medium: they can put emphasis on the quoted content (e.g. the message expressed by the quoted text, or the quoted post in the frame of the discussion) or on the quoted user (e.g. the relationship with the quoted user). Users make minute but meaningful editorial choices when deciding what to cite from a post and how to integrate the cited content in the quoting post[3]. Beyond single posts, quotes highlight the focal points in a discussion, and help maintain it on topic[4]. Quoting behavior captures social signals such as attribution, acknowledgement, and endorsement[5]. Quotes are an aspect of discussion that holds a wealth of information and yet has not been extensively investigated so far.

This work studies a forum's community through its implicit

[1] Enoch Peserico, Federica Bogo, and Vincenzo-Maria Cappelleri also contributed to this study, whose findings were first presented in Samory et al., "Quotes in forum.rpg.net", 2015; Samory et al., "Community structure and interaction dynamics through the lens of quotes", 2016; Samory et al., "Quotes Reveal Community Structure and Interaction Dynamics", 2017

[2] It might be no coincidence Twitter's recent introduction of "quote retweets" – a new feature allowing users to add their own comment to the verbatim copy of the retweet, encouraging discussion. Before the introduction of quote retweets, users employed workarounds to adapt retweets and replies to a wide range of use cases: Garimella et al., "Quote RTs on Twitter", 2016

[3] Niculae et al., "QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns", 2015

[4] Barcellini et al., "A socio-cognitive analysis of online design discussions in an Open Source Software community", 2008; Kang et al., "Analyzing answers in threaded discussions using a role-based information network", 2011

[5] boyd et al., "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter", 2010

network of interactions, rather than from its explicit social structure – this approach proved viable and effective in other occasions: Section **??** explains the advantages and caveats of this approach in finer detail.

The contributions of this study are threefold:

- *it provides insight on the role of quotes in discussion: while literature shows quotes support longer and more coherent discussions, this work gives evidence of how;*

- *it analyzes the implicit social network of four diverse forums: it proves that the quote network between users can reconstruct user identity and role;*

- *it gives a novel link prediction formulation that links the implicit and explicit social network: it proves that local characteristics of the quote network can predict friendship between users.*

The rest of the chapter is organized as follows. I start by taking a brief look at how quoting works, and how it differs from replying. Then, I focus on a number of basic quantitative metrics characterizing quotes in the four forums. Quote usage, albeit different in different forums, appears remarkably consistent across time and users in each forum. Also, although quotes share many of the typical traits of social interaction such as heavy-tailed distributions, they markedly lack "rich-get-richer" characteristics.

I then explore quotes in the context of the thread that surrounds them: one interesting finding is that quotes relay context between posts that are far apart in time, and help shorten threads efficiently. This suggests quotes play a crucial role in aiding thread navigation.

Next, I examine the implicit network that quotes effectively create between users. Using structural features of this network alone it is possible to re-identify a user across different discussions with fair accuracy. Also, PageRank[6] computed on the implicit quote network reveals core users in the forum communities better than the reputation mechanism embedded in the four forums. Moreover, is possible to predict if two users are friends with over 80% accuracy through local features of the implicit quote network.

[6] Brin et al., "The anatomy of a large-scale hypertextual Web search engine", 1998

Finally, I question what the implicit quote network may explain about community-wide phenomena: I review quoting patterns that are specific to each forum, and show how differences in these patterns correspond to differences in the type of community – as a case study, I identify defining characteristics that distinguish between forums providing advice by small groups of experts, and forums that are essentially large communities of peers.

I conclude with a discussion on the implications of these findings, in terms of security, interfaces, personalization, and community management, and I highlight opportunities for future research.

5.1 Research question

This research is driven by the following issues:

RQ1 Quotes are features of discussion. *Do quotes encourage discussion? In what ways does quoting facilitate discourse?*

RQ2 Quotes carry social signals. *Does social structure emerge between quote adopters? In what ways is this structure social? How does it relate to the ground-truth friend network in the forums?*

RQ3 Quotes reflect communities they are embedded in. *Can quotes characterize users, their relationships, and their roles? Do quotes provide metrics for comparing different communities?*

5.2 Quoting in online forums

Most online forums today offer a quotation mechanism, that allows a post author to cite excerpts of other posts – either in the same or in other discussion threads. To do so, one simply clicks on a "quote" button that appears on the post to be quoted. This brings the entire quoted post, highlighted and preceded by "Originally posted by <quoted author>", into the new post at the current text insertion point. The new post's author then can manually edit the quoted post, and typically does so to remove less relevant passages (Figure 5.1 shows an example quote).

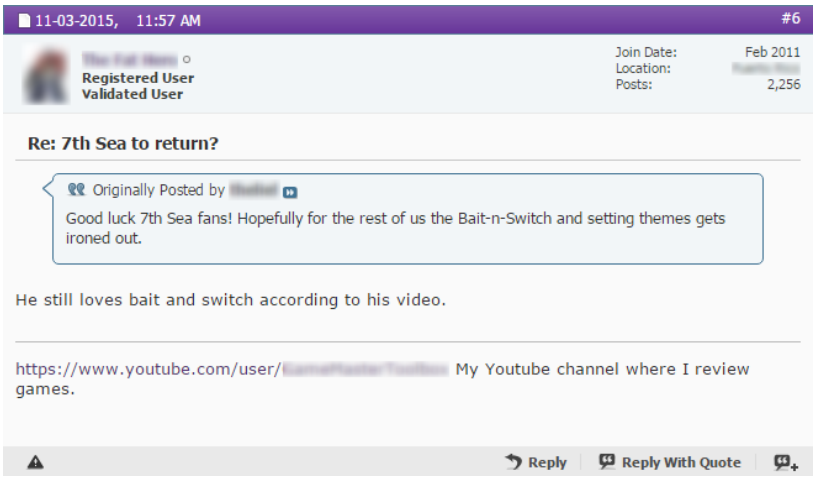


Figure 5.1: Example post containing a quote from RPG. At the bottom-right corner one can see the options for adding a new post: reply to this post, quote this post, and quote this post along with multiple other posts.

I remark that quotes are a widespread mechanism in forums, that differs from *replies* (analysed e.g. in Aumayr et al.[7]). A forum with replies links each post beyond the first to *exactly* one previous post *in the same thread* as a reply, effectively organizing the thread into a tree of posts rather than into a linear sequence. Unlike replies, quotes allow a post to link multiple previous posts (or none), potentially belonging to other threads or even subforums. Furthermore, quotes explicitly identify the portion of the linked post to which they refer, and users make minute but meaningful

[7] Aumayr et al., "Reconstruction of Threaded Conversations in Online Discussion Forums", 2011

editorial choices when deciding what to cite from a post and how to integrate the cited content in the quoting post.

5.3 *Quotes as metrics*

Before focusing this investigation on individual threads, I would highlight three aspects of quotes that are present in all four forums and strike as unusual: 1) quote usage that varies widely across different users and forums, but is markedly consistent on average within each forum over many generations of users and large fluctuations of post volume 2) heavy tails in the absence of "rich-get-richer" phenomena and 3) matching tails for quotes made and received.

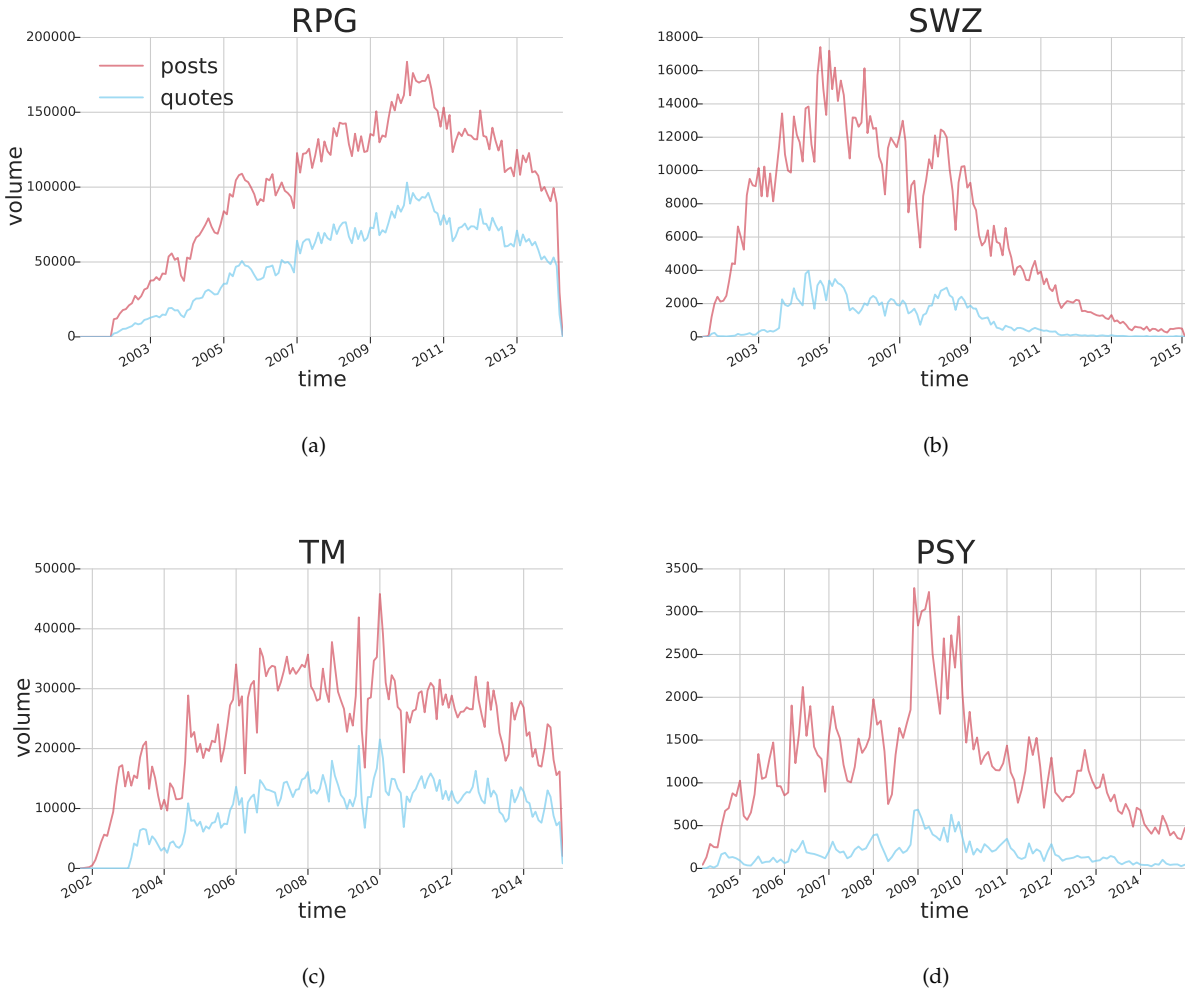


Figure 5.2: Quote and post volume per month. While the quotes/posts ratio varies across forums, within each forum it remains almost perfectly constant month after month. The ratio is higher in open discussion forums (RPG, TM) than in support forums (SWZ, PSY).

5.3.1 *Prevalence in the four forums*

Quote usage is widespread, but varies between different forums, with a ratio of quotes/posts ranging from $\approx 60\%$ in *RPG* to $\approx 20\%$ in *SWZ* (see Table 3.1 and Figure 5.2). Also, individual users exhibit a wide variability in terms of quotes they make and/or receive.

However, *within each forum the quotes/posts ratio remains almost perfectly constant over time*. This is particularly surprising given that not only does the post count change significantly from month to month, but that the average user "lifetime" (less than 1.5 years for all four forums) is significantly shorter than the time interval under observation. The quotes/posts ratio then appears to be an extremely specific signature of each forum's language and interaction patterns, suggesting the existence of an independent "kansei" of each forum that, although emerging from the behaviour of individual posters, assumes and actively maintains a relatively unchanging identity of its own by shaping the behaviour of subsequent generations of posters.

5.3.2 *Distribution across posts*

The power-law exponents of the quote distributions per post vary between forums, ranging from ≈ 3 in *PSY* to ≈ 4.5 in *RPG*; but within each forum the power-law exponent for quotes made *to* a post almost perfectly matches that for quotes made *by* a post (see Figure 5.3). This is true even at the extreme end of the spectrum, with the exception of a very few highly quoted posts in *RPG* (then again, a remarkable post in *RPG makes* no less than 79 quotes).

This may be surprising given that making a quote, as opposed to receiving one, requires some effort by the poster – and is indeed in contrast with what one observes in many other social contests marked by a similar effort asymmetry, from citation networks to the World Wide Web, where the largest number of citations/links/etc. received by a node typically far outstrips the largest number made[8].

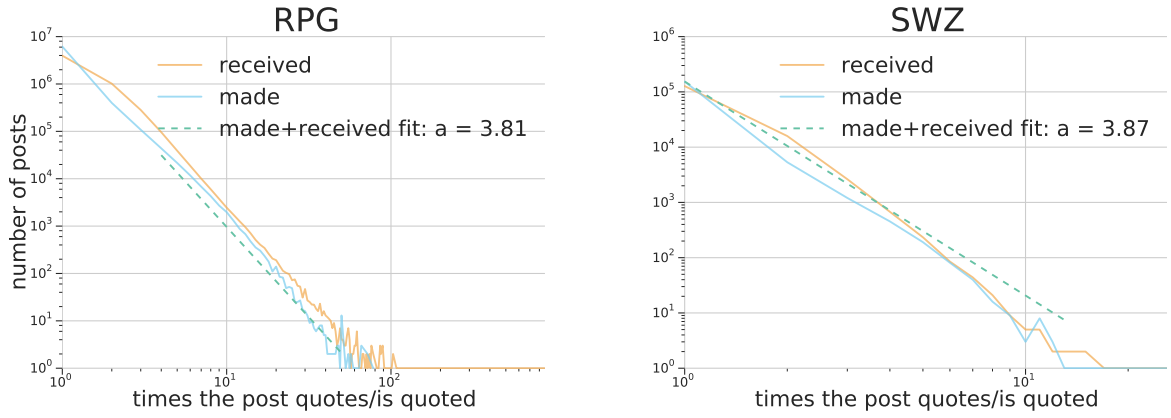
[8] Newman et al., "Why social networks are different from other types of networks", 2003

5.3.3 *Distribution across users*

In all four forums the distribution of quotes/post is heavy-tailed (see Figure 5.3). Heavy tails are a common phenomenon in the most diverse social settings, from co-authorship networks to salary distributions, and are typically explained through the so-called "rich-get-richer" effect. However, *none of the four forums sports rich-get-richer effects*: more prolific authors receive (and make) more quotes, but no more and no less than ensembles of less prolific authors with the same aggregate post count (see Figure 5.4). This is consistent with the fact that none of the forums makes visible, to users, how many quotes another user or post has received – but it leaves the observed heavy-tails without their "usual" explanation.

5.3.4 *Quotes as forum signatures*

Before moving on to analyzing activity on the forums through quotes, I wish to highlight that the preliminary analyses presented in this section give confidence on quotes as a measure unit of forum activity. The widespread adoption of quotes makes them strong, stable signals. Also, quotes show consistent patterns independently



(a) (b) (c) (d)

Figure 5.3: Distribution of quotes per post. The distribution is heavy-tailed; the best-fit power-law exponents are also reported.

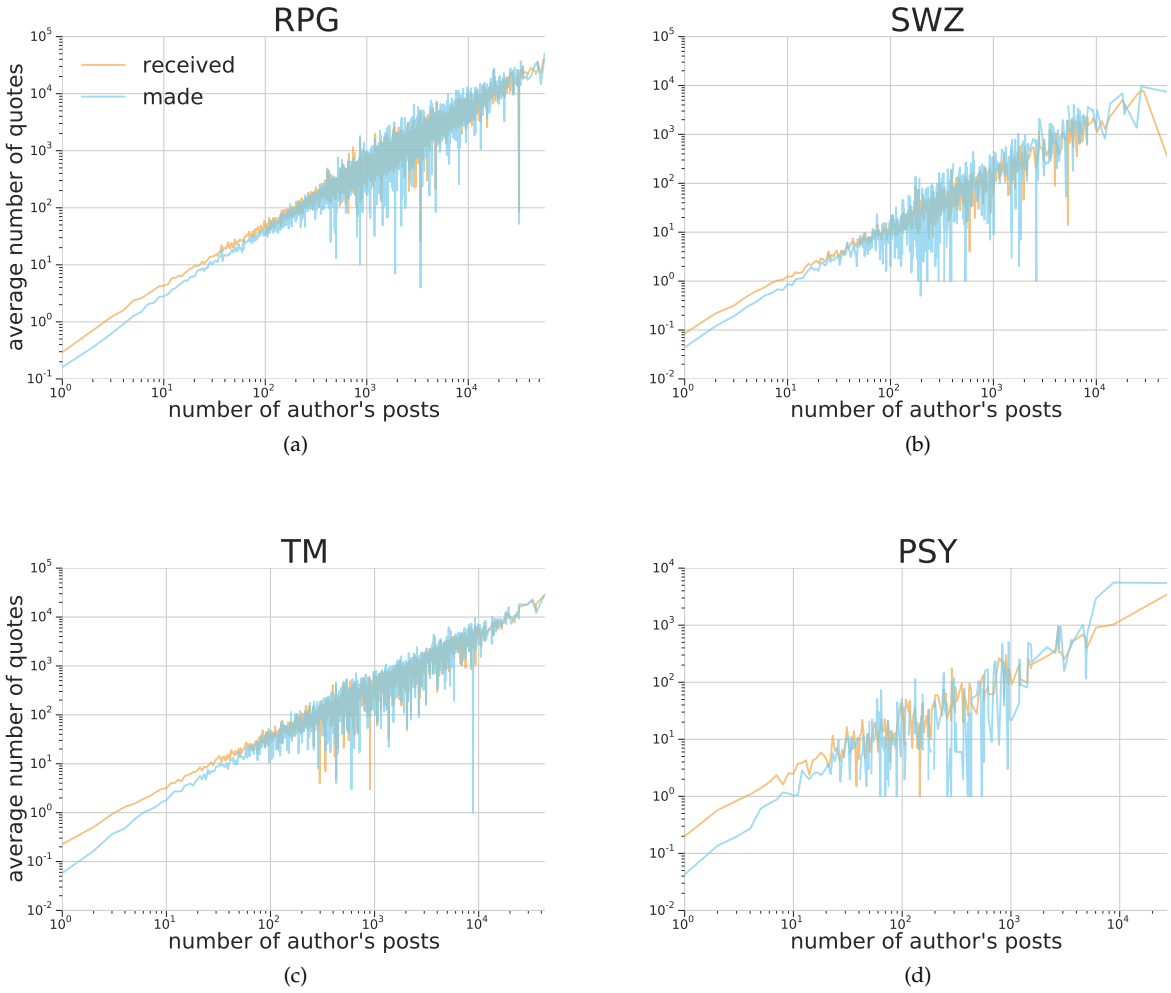


Figure 5.4: Average number of quotes vs. number of posts per user. Note the absence of rich-get-richer phenomena: users with high post counts make and receive as many quotes as ensembles of users with the same aggregate post count.

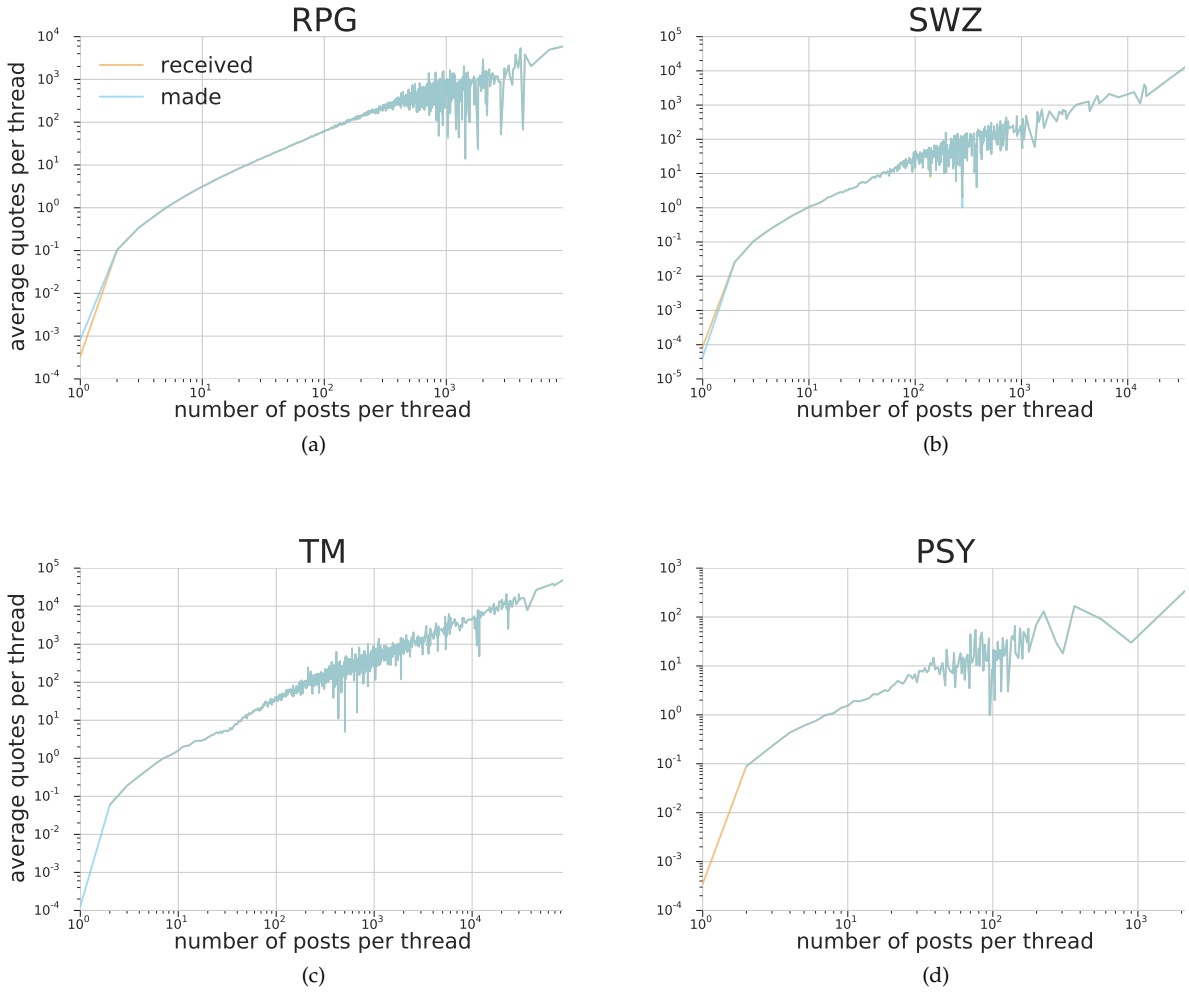


Figure 5.5: Average number of quotes vs. number of posts per thread. Longer threads sport a relatively higher fraction of quotes.

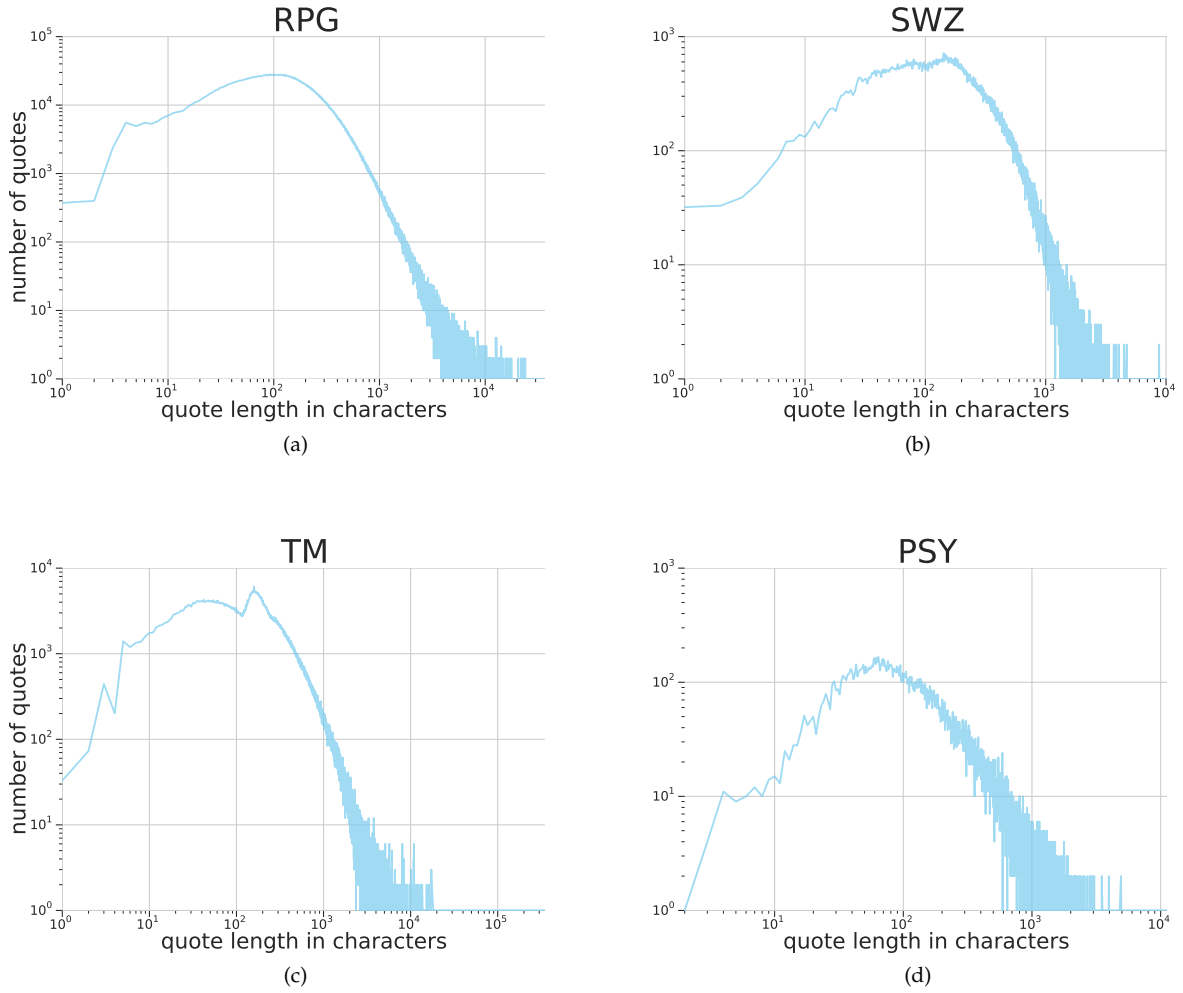


Figure 5.6: Quote length distribution (number of characters). Shorter quotes are comparatively rare, probably due to the difficulty of providing context in less than 140 characters; longer quotes exhibit a heavy-tailed length distribution.

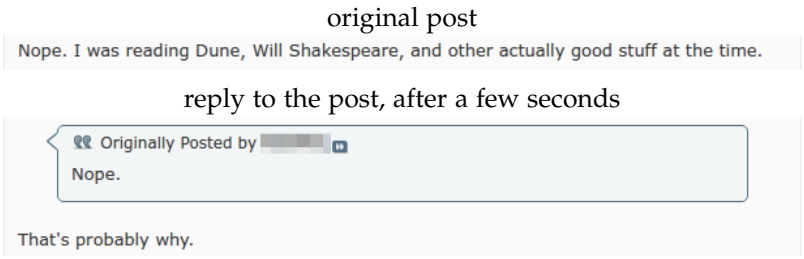


Figure 5.7: Example of a "short quote" with small quote delay. When a post quotes another that is close in time, all users involved are on the same page of the discussion, and quotes tend to be short and to the point.

of the deep differences in the forums studied, which bodes well for generalizability. Moreover, quote patterns are robust to the generational turnover in the forums. The simple relations between quotes and other well understood quantities, such as posts per users or the overall volume of posts, makes interpreting quote metrics easier. Nonetheless, quotes pick up complex and nuanced phenomena at microscopic as well as macroscopic scales. Quotes show all qualities needed by valid sources for metrics, which makes them preferable to other sources, like users' post count or friendship links, that are often highly varying, noisy, or sparse.

5.4 *Quotes and discussion*

Let me now zoom to the level of detail of individual threads, uncovering a unique role that the quote network plays at this scale – sustaining discussion and helping navigate long threads.

5.4.1 *Quotes provide context for asynchronous communication*

The previous section shows that the average number of quotes both made and received by authors with a given post count is almost perfectly proportional to that post count. The same cannot be said of quotes by/to a thread. Short threads both make and receive relatively fewer quotes per post (see Figure 5.5). A possible explanation is that *a unique role of quotes is to aid intra-thread navigation* – with shorter threads being intrinsically easier to navigate and thus requiring less quote support.

Analysis of quote length supports this hypothesis. Quote length follows a heavy-tailed distribution, at least beyond a minimum threshold of 140 characters[9] (see Figure 5.6) – shorter quotes are comparatively rarer, showing the difficulty of conveying meaningful information with a chunk of text shorter than a tweet. While a few of the very shortest quotes are essentially typing/posting errors, the majority of quotes of even 2 characters appear valid (e.g. "no", "3?", "me"); most of these tiny quotes refer to a very "close" post on which they rely to provide the appropriate context (see Figure 5.7). And indeed, quote length markedly grows with the temporal distance between quoting and quoted post (see Figure 5.9, and compare Figures 5.7 and 5.8).

[9] using the python module "powerlaw": arXiv:1305.0215

Figure 5.8: Example of a "long quote" with large quote delay. When a post quotes another that is distant in time the quote itself must provide the appropriate context.

5.4.2 *Quotes help navigate massive multiparty discussion*

Another way to observe this phenomenon is to consider the *depth* of posts, defined for the initial post of any thread as 0, and for any other post p as 1 plus the minimum depth of any post that p quotes or immediately follows in the thread – in some sense, the depth of a post being its distance from the opening post, in the thread *augmented by the quote network*. In the absence of quotes, both maximal and average post depth would be proportional to thread length. However, in practice, quotes provide shortcuts to the discussion, compacting longer threads more than short ones, in terms both of average and of maximal post depth (see Figure 5.10).

It is not entirely clear whether (forums with) longer threads tend to generate more quotes, or instead (forums whose culture generates) abundant quotes can more easily sustain longer threads. However, Figure 5.10 shows that if one compares threads *of the same length* from different forums, those from forums with greater averages of thread length and quote density (like *RPG* and *TM*) tend to have more quotes and lower depth than those from forums with lower averages (like *SWZ* and *PSY*). Thus, the simplest explanation is that quote abundance is an intrinsic characteristic of each forum (consistently with the findings of the previous section) that is the cause, rather than the effect, of longer discussions. As I will discuss in the latter part of this chapter, this has some practical implications on interface design and moderation policies.

5.5 *Quotes identify and characterize users*

For many years, forums have been the venue of choice for communities of users sharing interests on a topic. However, forum users can interact almost only through discussion – forums mostly lack the devices of modern online social platforms, such as friendship, liking, and reputation mechanisms. Even if quotes are indicators

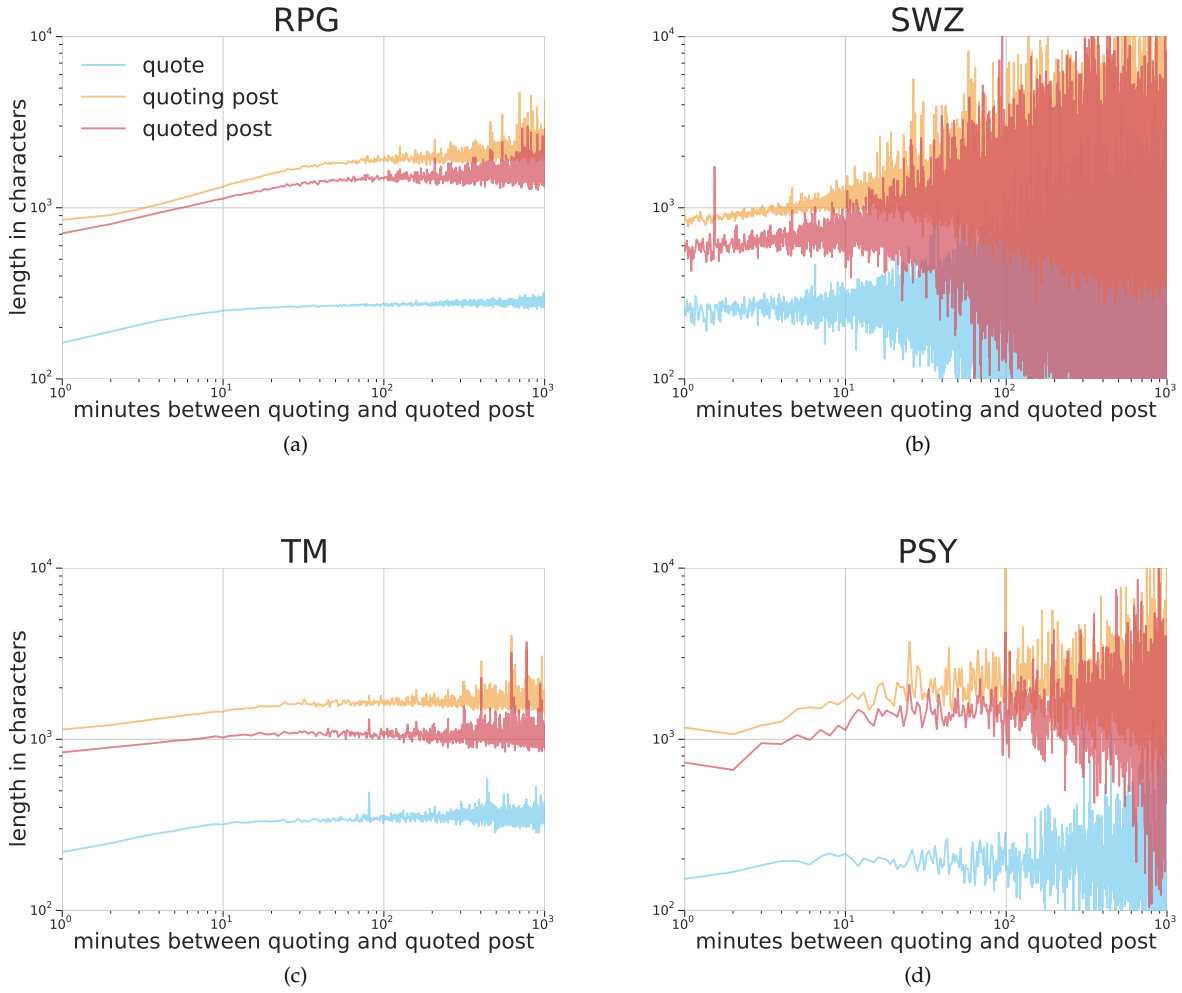
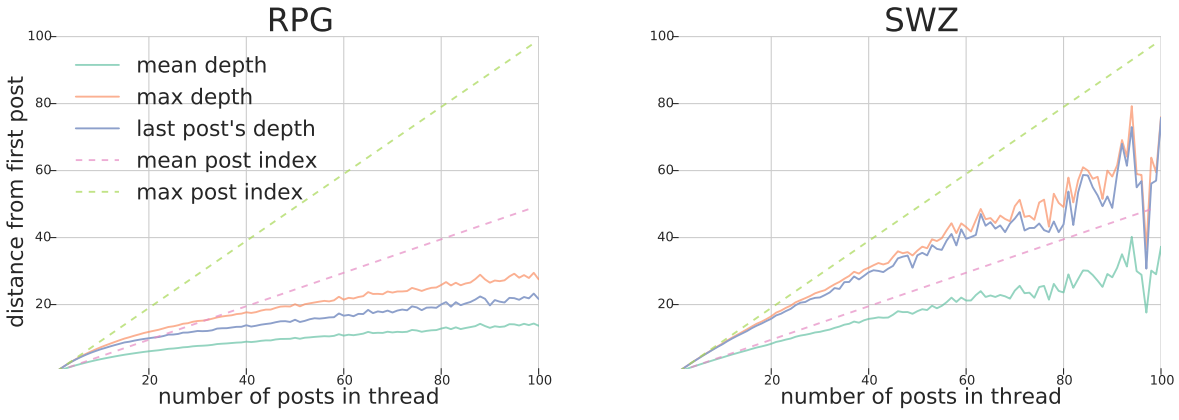


Figure 5.9: Quote delay vs. length of quoting post, of quoted post, and of quote. When quoting and quoted post are distant in time, quoted text tends to be longer, possibly because quotes must provide context.



(a) (b)

(c) (d)

Figure 5.10: Maximum, average, and last post's depth. Depth of a post is the shortest distance of that post to the opening post "along" quotes (assuming each post also implicitly quotes the post immediately above it). Post depths tend to decrease sublinearly with thread length, and they are smaller in forums with longer discussions (RPG, TM).

of attention, common interest, and attribution, forums do not tally them: there is no immediate way to learn which or even how many users have quoted a given user or post. In this light, it may be surprising that it is possible to retrieve a latent structure of a forum's community by observing how users quote one another, and that this structure shows the typical features of a social network.

5.5.1 *Modeling the quote network*

The rest of this chapter analyzes quotes between users, instead of between posts. To build this implicit social network (see Section **??** for related research), I simply collapse all posts authored by the same user into a node corresponding to that user. The resulting *author* quote network can therefore be defined as the directed, weighted graph that has users as nodes, has an arc from user a to user b if a has ever quoted one of b's posts, with weight equal to the total number of times a has quoted b. As Table 5.1 shows, the author quote networks obtained from the four forums sport many characteristics of social networks.

5.5.2 *Analysis of the quote network*

metric	**RPG**	**SWZ**	**TM**	**PSY**
Nodes	35118	11544	9661	1553
Edges	2.5M	50.8K	291.7K	5983
Zero InDeg Nodes	3330	1832	533	117
Zero OutDeg Nodes	9628	6084	2996	853
NonZero Deg Nodes	22.2K	3628	6132	583
Unique directed edges	2.5M	50.8K	291.7K	5983
Unique undirected edges	1.8M	41.5K	203.9K	4804
Self Edges	4174	766	1265	70
BiDir Edges	1.4M	19.3K	176.8K	2.4K
Closed triangles	176.8M	250K	5.4M	11.4K
Open triangles	1.4G	11.9M	51.2M	715K
Frac. of closed triads	0.111	0.021	0.0962	0.0158
Conn. comp. size	0.995	0.975	0.993	0.993
Strong conn. comp. size	0.625	0.283	0.625	0.365
Approx. full diameter	7	7	7	6
90% effective diameter	3.317	3.690	3.338	2.922
Average clustering	0.385	0.305	0.469	0.431
Assortative mixing	0.088	-0.249	0.226	-0.004

Table 5.1: Author quote network statistics show several characteristics of social networks, such as sparsity, low diameter, high clustering coefficient.

First, the author quote networks of all four forums are sparse, containing only a small fraction of all potential edges. Second, they are small worlds, with a giant connected component. More precisely, all forums show a weakly connected component that includes more than 97% of all nodes. The strongly connected components include approximately 62% of all nodes in the case of *RPG* and *TM*. These numbers closely match the corresponding values,

92% and 68% respectively, for the Twitter network [10] (the strongly connected component of *SWZ* and *PSY* is slighly smaller, around 30% – but see below). Furthermore, the diameters for the largest components are relatively small: the approximate diameter is 7, and 90% of all nodes are within 4 hops of each other despite the graph's sparsity.

[10] Myers et al., "Information network or social network? The Structure of the Twitter Follow Graph", 2014

Also, quotes are highly reciprocated: roughly 50% of all node pairs connected by an arc sport an arc in the opposite direction, and $2 - 10\%$ of all triads are closed. The clustering coefficient, too, is remarkably high (above 0.3); in particular, it remains high even for nodes of high degree, definitely more than in the Twitter or Facebook graphs [11] – a possible explanation lying in the highly specialized nature of forums that tends to limit the variety of a user's circles.

[11] Myers et al., 2014

Finally, assortativity by node degree (informally, the propensity of nodes to link to nodes with roughly the same degree) is mildly positive for *RPG* and *TM* (like in the Facebook or Twitter graphs[12]), and mildly negative for *PSY* and *SWZ* (as the Internet and WWW graphs[13]). This finding is somewhat surprising, considering the lack of *rich-get-richer* phenomena for users with high post count. An explanation might be that quoting follows social conventions different from simple posting and replying. The differences between *RPG* and *TM*, and *SWZ* and *PSY* match the intuition of the first pair of forums being driven by more "social", peer-to-peer conversations, and the second pair of forums being venues for obtaining information from experts.

[12] Myers et al., 2014

[13] Newman et al., 2003

5.5.3 *Quotes identify users*

Quotes provide a wealth of information about individual authors. As a very first step in this direction, I show that quoting patterns of individual users are in some sense weak digital fingerprints: if one takes a set of users, and partition their posts into two groups, it is possible to match users in the two partitions comparing the author quote networks built within each partition.

More precisely, I take a randomly chosen group of n users, $n \in [2, 5, 10, 20, 50]$ (I only consider users with at least 100 posts, to remove noise). For each user in the group, I partition each of the threads he appears in, so that the total number of his posts in each partition is approximately balanced. Then, for each partition, I build the corresponding author quote network, using all quotes received from and made to the posts in the partition – taking care to remove posts present in the other partition, if any (recall that author quote networks are directed graphs where users are nodes and arcs are quote links between them, weighted by the actual number of quotes). For each of the n users and for both graphs, I compute several network metrics characterizing the user's ego network. The resulting feature vectors are then $L1$-normalized, after replacing missing values with the average value for the respective feature. I

correctly identify a user if his feature vectors in the two partitions are the closest in terms of cosine similarity. I evaluate the identification algorithm using accuracy. I repeat the process 10 times per forum, to stabilize results. The complete list of the network metrics taken into consideration can be found in the appendix.

Accuracy values exceed 80% in all cases when attempting to discriminate between two users, and decrease to around 30% on average for 50 users.

The accuracy achieved consistently and considerably surpass the random baseline in all four forums. Accuracy is also comparable to other approaches from the authorship attribution literature, where identification is performed analysing the text of user posts (see results in the). However, note that *I completely ignore text*, and instead only make use of the quote network.

#users	RPG	SWZ	TM	PSY
2	1.00	0.85	0.80	0.85
5	0.80	0.50	0.58	0.58
10	0.69	0.49	0.57	0.48
20	0.54	0.39	0.45	0.36
50	0.41	0.24	0.28	0.24

Table 5.2: Accuracy for user identification via the quote network.

5.5.4 *Quotes expose prominent users*

Forums often present each user's post count next to the username, as a form of status badge. However, it is often the case that some prominent members of a forum – from moderators to "experts" of the field – are not very active posters, and thus have a relatively low post count. It turns out that the quote network can identify these prominent users.

PageRank[14] is probably the most famous algorithm to identify "important" nodes in a graph. Originally proposed to rank Web pages, it is employed in an ever-growing number of very diverse fields ranging from spam detection and word sense disambiguation to gene ranking. Informally, the PageRank score of a node v in a graph is the stationary probability of a "random surfer" being on v at any given time, if that surfer starts and "occasionally" restarts from a random node of the graph, and at all other times follows at random one of the outgoing arcs of the node that the surfer is currently on.

[14] Brin et al., 1998

For each of the four forums, I report in Table 5.3 the PageRank scores of all nodes in the *unweighted author quote network* (where each node corresponds to an author, and an arc connects two authors if the first has quoted the second at least once), using the "typical" damping score 0.85 (see Brin et al.).

PageRank identified many prominent users of each of the four forums under examination: most of the authors of high PageRank appear to either have some official position (e.g. Moderator) in their

respective forum, or have an "important" role even though it may not be immediately evident to a casual onlooker. One example is a user from *RPG* ($PR = 5$ in Table 5.3) who was one of the players at the gaming table of the late Gary Gygax (the author of Dungeons and Dragons, the very first and still best-selling roleplaying game). Another example is a pair of *TM* users ($PR = 1\&5$) who are editors of the review website associated to the forum. Conversely, the PageRank score of a "dummy" author in *SWZ* (used for trivial maintenance tasks on the forum) is very low, despite its post count being the highest.

Similar results can be obtained computing the PageRank scores on the forums' *weighted* author quote networks (where each arc is assigned a weight, and a selection probability, proportional to the number of quotes from one author to the other). In particular, in each forum the majority of the 20 authors with the highest PageRank scores on the unweighted author quote network is also among the 20 authors with the highest scores on the weighted network, and vice versa.

5.6 *Quotes reconstruct friendship links*

The previous section showed that quote networks can reveal information about the identity and role of authors. It is natural to ask if they can also reveal *ties between* authors – e.g. friendships. To answer this question with some rigour, ground truth data is necessary to compare information obtained from the quote network. Fortunately, the forums I analysed can provide such data through their little-used friendship mechanism.

5.6.1 *Analysis of the friendship network*

The vBulletin platform allows a (registered) user to visit another (registered) user's profile, and send a friendship request; if the recipient accepts the request, the two users will thereafter appear in each other's friends list. The friendship mechanism sees little use in all four forums (less than 10% of all users), presumably because of its limited integration with the other forum services. Analyzing the friendship network (statistics are reported in Table 5.4), one can see that most users who have at least one friend are connected through one or more degrees of separation (more than 90% of nodes using friendship mechanisms are within a giant connected component, and more than 80% of friendship edges are between nodes of that component). However, both the average degree of the friendship graph and the number of closed triangles in it are quite low, contrary to typical social data. *The forums' friendship networks are therefore very sparse, and can be interpreted as a noisy subsample of the underlying community structure.*

PR	#Posts	Score	RPG Role	#Posts	Score	TM Role	#Posts	Score	SWZ Role	#Posts	Score	PSY Role
1	41381	9	Reviewer	18302	16	Reviewer*	24683	4	Mod	27212	1	Founder
2	56543	1		16590	20	Admin	27407	3		6048	3	Admin
3	48440	5		43025	1		28934	2	Mod	8803	2	Mod
4	35028	12		10881	51		19066	5	Mod	4588	7	
5	31205	23	VIP	15094	26	Reviewer*	10188	11		4847	6	
6	34541	15	Admin	18239	17		17972	7	Mod	4883	5	
7	18587	81		11610	45		18826	6	Mod	2565	14	Mod
8	24452	42		22793	10		5933	32	Mod	5036	4	
9	34298	16		13649	37	Reviewer	3582	63	Mod	3610	9	
10	23908	43		23477	9		11317	10		3091	10	
11	24759	38		20593	12		8310	15	Mod	2747	12	
12	20454	66		14185	32		5748	35		1535	17	
13	38357	10	Mod	11243	48		7175	23	Mod	3622	8	Admin
14	14096	149		8954	81		13613	88		833	34	
15	24502	40		13437	38		9917	12	Founder	871	31	Banned
16	48377	6		10516	54		3889	56		2861	11	
17	45908	8		10922	50		7273	22		1551	16	
18	32879	20		10437	57		6036	31	Mod	2137	15	
19	18955	77	Banned	9227	74		3705	60		873	30	
20	33114	18	Mod	11641	44		2614	83		1023	24	Admin

Table 5.3: Member total post count, post count rank, and social role for the 20 members with highest PageRank, in each forum. "Reviewer*" tags users who are reviewers on the site, but lack explicit indication of the fact in their member pages.

metric	RPG	SWZ	TM	PSY
Number of nodes	3920	112	927	136
Number of edges	8040	232	2929	177
Average degree	4.10	4.14	6.32	2.60
Connected components	245	1	1	12
Frac. nodes in largest cc	0.85	1.0	1.0	0.84
Frac. edges in largest cc	0.95	1.0	1.0	0.94
Diameter of largest cc	13	4	4	10
Closed triangles	4607	0	0	30
Open triangles	311335	3296	440410	1407

Table 5.4: Statistics for the friend network. In all four forums only a minority of users adopt the friendship mechanism.

5.6.2 *Quotes between friends*

The next section reconstructs friendship links from features of the quote network. Before proceeding to the analysis, however, I want to remark that the simplistic approach of predicting as friends pairs of users that quote each other frequently would not work. Figure 5.11 shows the distribution of how many quotes a pair of users share, depending on whether they adopt the friendship system or not. I consider the cases where no user in the pair uses the friendship system; only one does, both do, but the two are not friends with each other; and the two are friends. It is true that users that adopt the friendship system quote each other relatively more frequently (the tail of the distribution is fatter). However, the range of the number of quotes exchanged within each pair is similar in all four conditions, and the overall volume of quotes exchanged by users using the friendship system is lower, making a threshold on the number of quotes a problematic predictor. Most strikingly, friends in the SWZ and TM forums *never* quote each other directly (to be exact, there are two quotes between friends in TM: being a single data point, no line shows in the plot). The friendship network and the quote network are two very distinct graphs, and it is not straightforward to infer properties of one from the other.

5.6.3 *Reconstructing friendship from quotes*

Since most users do not *not* use the forum's friendship mechanism, I simply gauge to what extent quote data can reconstruct the friendship lists of those users who do. It is reasonable to assume that the same method would predict with similar accuracy "invisible" ties between users eschewing the forums' friendship mechanism.

For each forum, I randomly sample 50 pairs of forum members who have explicitly acknowledged each other as "friend" using the forum's friendship mechanism, and 50 pairs of members who have used the friendship system, but have not befriended one another. In particular, I obtain the latter pairs sampling 50 users with a probability proportional to their degree in the friendship graph, and sample a number of users who are not their friends, with probabilities

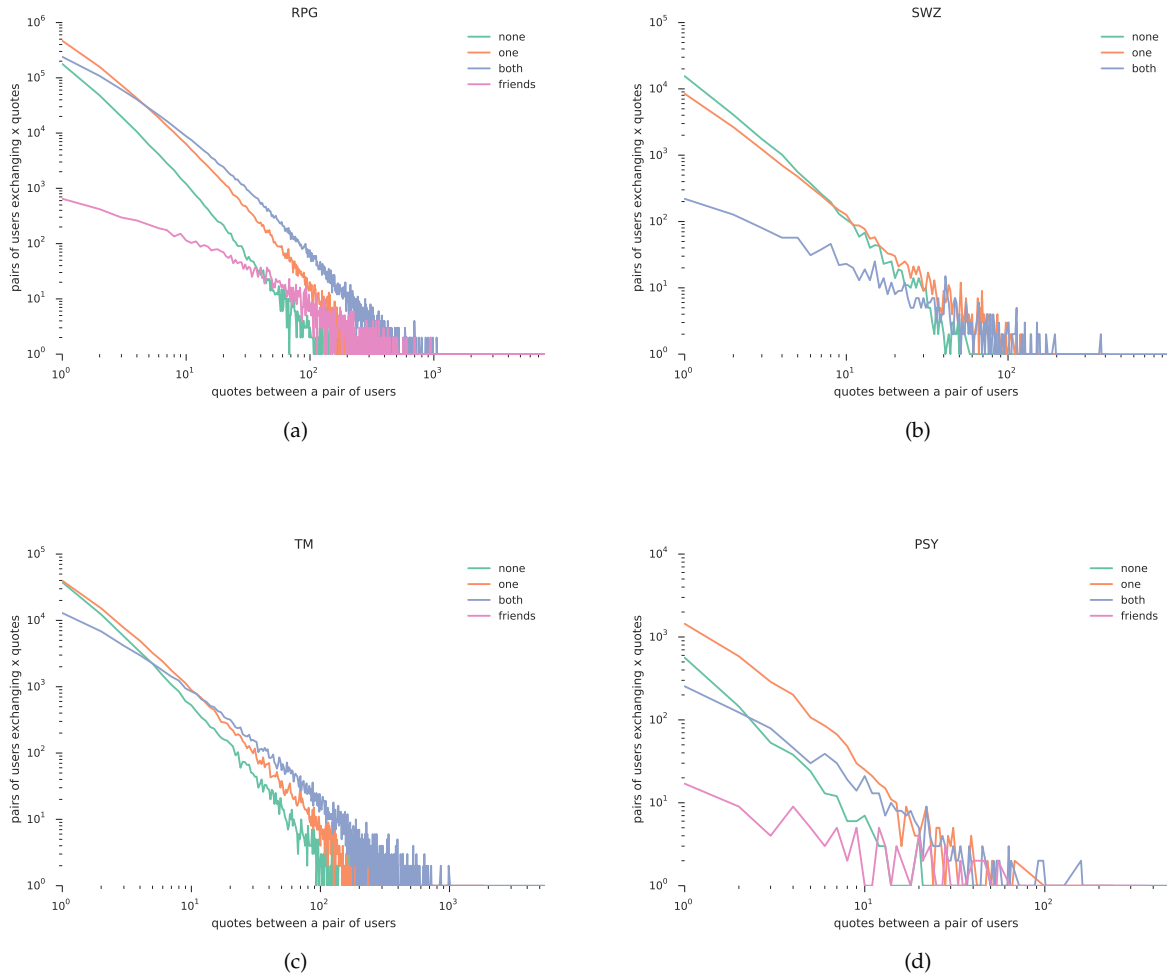


Figure 5.11: Quote multiplicity distribution for pairs of users with different degrees of adoption of the friendship system: no user in the pair uses the friendship system, only one does, both do but the two are not friends, the two are friends. The higher the adoption, the flatter the distribution: friend users exchange more quotes with each other. However, overall, friend users in SWZ and TM do not share quotes at all.

again equal to their respective degree. I then build feature vectors using network metrics that are local to the nodes and their ego network, as well as metrics of co-occurrence in threads. The metrics used are listed in the appendix. Finally, I evaluate the accuracy of a Logistic Regression classifier in 10 rounds of random partitioning into train and test set, maintaining an $80-20$ proportion and class balance.

Accuracy appears on average around 70%. If one frames the problem at a more local scale, predicting if two users posting *in the same thread* are friends or not, the average accuracy rises above 80%. Accuracy results for all datasets are reported in Table 5.5.

sampling	**RPG**	**SWZ**	**TM**	**PSY**
degree-based	0.755	0.730	0.660	0.670
thread-based	0.730	0.890	0.885	0.715

Table 5.5: Accuracy in friend prediction.

feature	RPG	TM	SWZ	PSY
#edges				-0.36
ego_acc	0.02	** 0.23	*** -0.68	0.01
#edg_btwn_friends	*** 1.34			*** -1.55
reciprocity			0.05	0.15
weight_reciprocity	*** -0.32	0.19		*** -0.60
exclusivity	** 0.91			*** 1.25
mixin_ii	-0.03	-0.12	0.02	-0.08
mixin_io	* 0.19	*** -0.41	** 0.23	** 0.29
mixin_oi	-0.07	*** 0.46	-0.09	** -0.27
mixin_oo	-0.10	-0.15	0.15	*** 0.30
nmin				-0.30
avg_neighb_dmin	*** 0.59		*** -1.00	*** -0.30
avg_neighb_dmax	* -0.19	** 0.15	*** 0.27	-0.06
jaccard	*** -0.35	*** -0.98	0.07	-0.12
preferential_attach	*** -2.39		-0.36	
jaccard_thread	*** 2.51		-0.09	*** 1.64
#common_friends			-0.06	
pref_attach				-37.97
directionality	-0.015			
dispersion_b	0.11			
nmax	0.12			
delta	*** 1.74			

Table 5.6: Feature β weights obtained in the Logistic Regression model for friend prediction. Asterisks represent statistical significance: * p < 0.1, ** p < 0.05, *** p < 0.01.

5.6.4 *Feature performance*

It is possible to gain more insight into which features best predict if two users of the forums are friends through the β coefficient that the Logistic Regression model assigns to each feature. Since features are potentially correlated, I perform two additional pre-processing steps before computing the β coefficients: whitening to avoid disproportionate feature scales, and feature selection through Randomized Logistic Regression to avoid multicollinearity. β coefficients for the four forums are reported in Table 5.6.

Several features maintain the same role across all four forums. Exclusive relationship between two users is intuitively a positive predictor of friendship. Other positive predictors are the Delta measure and Jaccard index computed on participants of the same threads. Jaccard index and preferential attachment computed on the author quote network are negative predictors – consistently with the strong positive influence of relationship exclusivity. Weighted reciprocity (measuring how reciprocated are quotes between two users, weighted by the total number of quotes by each user), has a negative coefficient, somewhat counterintuitively. Overall, small groups of users that mostly exchange quotes among themselves, and that appear in discussions sharing similar participants, are more likely to be friends in the forum.

Some features exhibit different behavior in different forums.

This is may be explained by the different goals of users in different forums. For example, the β coefficients for the number of arcs between users in the intersection of the quote ego networks is negative in *PSY* but positive in *RPG*. This may be because users in *PSY* come to the forum to address delicate, highly personal problems, making isolated friendships preferable to tightly knit circles, in order to ease social pressure. Assortative mixing of input and output degrees, as well as average neighbor degrees and average clustering coefficients, sport β coefficients with inconsistent signs – however their absolute value is small, indicating a relatively small impact on friendship prediction.

5.6.5 Quotes as signals of friendship

Summing up, how many quotes two users exchange is not directly indicative of whether they are friends or not. However, the structure of the "quote neighborhoods" of the two users – how well connected are the users they quote – can reconstruct user friendship with over 70% accuracy, albeit the intrinsic noise in the ground truth on friendship. Quotes capture friendship signals in a subtle way: this signal is mediated by the network of users the two friends are mutually connected to.

5.7 Quotes characterize communities

I began this inquiry analysing how quotes link, within each thread, individual posts. I then "zoomed out", looking at how quotes link, within each forum, individual users – revealing their identities, ties, and prominence within the forum. I now "zoom out" one step further, and look at how quote networks of entire forums compare to each other and yield insights on each forum's community as a whole.

To this end, I partition the four forums into two pairs. The first pair includes *SWZ* and *PSY* – two forums (the first in Italian and the second in English) sporting an "elite" group of expert users, that provide advice and support to a larger group of less experienced users seeking help (on software and mental problems, respectively). The second pair of forums includes *TM* and *RPG* – two forums (again, the first in Italian and the second in English) where the relationship between users is much more one between peers discussing a favourite subject (music and roleplaying games, respectively), even though users might range from casual hobbyists to experienced professionals. The rest of this section shows how differences in the quote networks, as highlighted in the previous analyses, may reflects in this distinction between "elite-" and "peer-discussion" forums .

5.7.1 *Engagement in discussion*

One first axis of distinction between the two pairs of forums is the type of discussions users engage in: *SWZ* and *PSY* are venues for question answering, troubleshooting, and gathering expert opinion; *TM* and *RPG*, instead, are driven by peer debate, reaction to events, and comments on themes of shared interest. One simple metric of engagement in discussion is the overall *quotes-to-posts ratio* – the higher the ratio, the higher the engagement, since each post prompts more reaction. Unsurprisingly, the quotes-to-posts ratio is higher in *TM* and *RPG* – which have more numerous, longer discussions between users; in *SWZ* and *PSY* instead, discussions are shorter, and individual users post less frequently. A more sophisticated metric would be the *last post's depth distribution* – informally, the minimum length of a thread if one can "shortcut" through quotes. One can expect users more engaged by the whole content of a thread to create more far-ranging quotes, reducing depth. And indeed, threads in *TM* and *RPG* are longer on average, but also sport comparatively smaller depths.

5.7.2 *Community connectivity*

With the exception of the minority of experts, the average user in *SWZ* and *PSY* is likely less involved in the communal life, coming to the forum only when in need of advice on a specific problem. This is reflected in quote metrics. The *fraction of users in the strongly connected component* – the group of users that can all reach each other through a sequence of quotes – is smaller in *SWZ* and *PSY*, revealing a smaller core of users, presumably built around experts. The *fraction of closed triads* adds further evidence. Triadic closure, the tendency to befriend a friend of a friend (in this case, to share a quoting/quoted user with another user one quoted/was quoted by), is significantly lower in *SWZ* and *PSY* than in *TM* and *RPG*. The customary explanation for triadic closure in social networks is cognitive balance: if A and B each have strong ties to X, lack of ties between A and B would put X in a "socially unstable" position.

From this point of view, a lower fraction of closed triads in the author quote networks of *SWZ* and *PSY* suggests the two forums are less tightly knit on average (or have a smaller group of users that are tightly interconnected) than *TM* and *RPG*. This is coherent with the fact that most *SWZ* and *PSY* users come to the forums seeking specific advice, and are less likely to build relationships with users in unrelated topics.

5.7.3 *Power differentials*

As I stated, relationships between users in the two pairs of forums are different, with users in *SWZ* and *PSY* showing larger power differentials by their very nature. One way to estimate power differentials in the user base is asymmetry in connectivity[15]. The smaller

[15] Gilbert, "Predicting tie strength in a new medium", 2012

fraction of closed triads in *SWZ* and *PSY* shows resistance in going beyond dyadic quote exchange. Even for pairs of users the *fraction of reciprocated quotes* is lower in *SWZ* and *PSY*. In fact, *assortative mixing by degree* – the propensity of users to connect to other users with similar number of quote links – shows that, in *SWZ* and *PSY*, casual users connect through quotes preferentially to users with high quote degree (typically experts, forum staff) and vice versa, while in *TM* and *RPG* quotes mostly link users with similar quoting patterns. The greater power differentials in *SWZ* and *PSY* are also reflected in the *distribution of PageRank scores* in the author quote network – which shows a sudden drop in *SWZ* and *PSY*, indicating a minority of users having disproportionate weight. Also, forum staff (e.g. administrators and moderators) is more represented in *SWZ* and *PSY* among the users with high PageRank scores, reinforcing the intuition of a larger power differential.

5.8 Discussion

Quotes in online forums are apparently simple tools, that nonetheless serve a variety of roles (from aids for intra-thread navigation to signals of common interest and acknowledgement) with finer-grained control than mentions, retweets and shares. And users appear to exploit them to the fullest – indeed beyond their original design, suggesting the possibility of additional uses like quote-based recommendation or "soft" moderation.

Quotes thus yield a wealth of information about both individual users and the community they are part of. In particular, this work showed how quotes in online forums can be used to "fingerprint" authors and communities, to identify prominent authors, and to expose otherwise hidden friendship relationships – with double-edged implications in terms of privacy, security and personalization.

5.9 Implications

The number of different roles that quote play, and the wealth of information they yield (at all scales of the discussion – from posts in threads, to individual authors in forums, to entire communities) have a number of implications for users, community managers, and interface designers.

5.9.1 Theoretical implications

This work showed several results that complement existing literature.

First, literature showed that quotes help maintaining thread coherence, and highlighting central passages in a discussion[16]. This work suggests *how* quotes may do so: they carry more contextual information when referencing posts distant in time, and prove effi-

[16] Barcellini et al., 2008; Kang et al., 2011

cient shortcuts in the sequence of posts when trying to summarize an existing thread.

Moreover, literature attempted reconstructing links in the explicit social network from the implicit social network – Chapter 2 gives an overview of this line of research. However it either built the implicit network from independent activities of the users (e.g. two users are connected if they were both present at an event, or if both listened to the same album), or investigated cases where the implicit network is an overlay of the explicit (e.g. where only friends can interact). This work instead predicted links on the explicit social network from the topology of the implicit network. This framing is arguably preferable, in that it does not depend on arbitrary definitions of what constitutes a link[17], and does not assume any prior knowledge on the explicit network.

[17] How one defines a link results in structurally different topology: see De Choudhury et al., "Inferring relevant social networks from interpersonal communication", 2010

Finally, this work showed that quoting patterns show conventional adoption. This may impact the way research studies online communities. Social computing often addresses studying online communities in a top-down approach: fixing a real-world signal, such as status, and looking for correlates in the virtual domain. However, this approach is prone to confounding the target with other signals. This work suggests that interaction patterns allow an alternative, bottom-up approach: they can help study online communities through measurable, unambiguous features of online discussion, that can later be associated with (possibly multiple) real-world signals. If there are suspect confounders, they are easy to disambiguate through directly incorporating them in the model.

5.9.2 Practical implications

5.9.2.1 Implications for privacy and security

Users quote according to different patterns, and indeed quotes can be used to identify users. In fact, I have shown that quote interaction patterns can on their own provide an effective fingerprinting mechanism. This poses a serious threat to privacy in openly accessible communities. For example, quotes could be used to deanonymize available discussion datasets, with methods analogous to those employed to expose users in the Netflix prize[18].

[18] Narayanan et al., "Robust Deanonymization of Large Sparse Datasets", 2008

On the other hand, the highly characteristic quoting behaviour of individuals could be used to (help) detect compromised users accounts. A continuous authentication framework could track quotes, and notify moderators when these diverge excessively from the quoting patterns of each user. The same framework could provide (additional) evidence in connecting multiple accounts belonging to a single individual.

5.9.2.2 Implications for visualization

I showed that the quote network effectively "shortens" long discussion threads. In this sense quotes could be leveraged to develop

novel interfaces for thread navigation. For example, one could imagine automatically disentangling complex threads that involve multiple ongoing discussions into their individual components, based on the quote trellis. Alternatively, a visualization system collapsing all but the most quoted posts could provide an effective system of thread summarization. In fact, a similar approach based on replies (rather than quotes) has apparently proved effective [19].

[19] Kang et al., 2011

Also, quotes provide context that keeps conversation on topic. My findings suggest that the amount of context strongly depends on how far back in time the quoted post is – quotes to relatively recent posts rarely involve passages longer than one sentence. In these cases current interfaces, that involve quoting an entire post and then removing the majority of text as unnecessary, seem cumbersome. Interfaces could be considerably enhanced and streamlined by allowing the user to select the text to quote through a selection cursor, or by clicking on a version of the text that has been pre-tokenized – perhaps in an adaptive fashion, expanding with clicks from a single word (quick quote), to a sentence, a paragraph, and an entire post (extensive quote).

5.9.2.3 Implications for personalization

Different kinds of discussion communities may implement different personalization mechanisms through quotes. In Q/A, troubleshooting, and expert opinion forums, a few advanced users have a prominent role in the discussion. These communities use quotes in comment-reply patterns between advanced users and regular users (as I have shown, quoting is mainly dyadic and quote assortativity negative). Therefore, notifying users when they are quoted back would facilitate and speed up discussion. Communities that focus on peer discussion, instead, use quotes in a social-like fashion (higher fraction of closed triads, positive assortativity). In this context, users may want to be notified of new quotes made, by users they have themselves quoted, to other users.

I showed that quote networks can accurately reconstruct "hidden" friendship links between users. Platforms with friendship mechanisms can then take advantage of quotes to evaluate the de facto strength of such ties[20]. By the same token, quotes could be used as a recommendation system, to suggest new users one should follow in discussion platforms.

[20] Gilbert et al., "Predicting tie strength with social media", 2009; Roth et al., "Suggesting Friends Using the Implicit Social Graph", 2010

The "implicitness" of quotes, in contrast to explicit friendship links, may in fact be a desirable feature. Recommender systems are plagued by the filter-bubble effect, which is ultimately due to selective visibility of content and users. The quote network can substitute the friendship network to extract user features for hybrid recommenders (e.g. for new threads to read). The filter-bubble effect would be mitigated in quote-based recommenders, since quotes point to content that is interesting to the user independently of the content's author, quote links from said author are opaque to

the user, and prolific authors do not have disproportionate visibility (as shown, quotes do not exhibit the rich-get-richer phenomenon).

5.9.2.4 Implications for community management

PageRank score computed on the quote network provides a reputation score for users alternative to post count, that proved effective in identifying prominent users. Reputation based on post count is prone to promoting user profiles that are high on quantity but low on quality, like bots used for forum maintenance, or spammers. Quote PageRank not only penalizes low quality users, but also promotes reputable users independently of their post count.

Furthermore, I showed that forums sporting greater use of quotes also tend to sustain longer discussions; promoting quotes may thus increase overall engagement in discussion. Although causality and generalizability beyond forums should be investigated, early work on the recent introduction of quotes in Twitter supports this view[21].

[21] Garimella et al., 2016

Conversely, note that the best countermeasure to flamers and trolls is to avoid "feeding" them, i.e. to ignore their posts so as to deny them attention – in fact, community feedback and harsh moderation seem to exacerbate antisocial users' behavior[22]. If lack of quotes promotes thread death, then inhibiting the ability to quote posts, users, or threads could be an effective, lightweight moderation tool.

[22] Cheng et al., "Antisocial Behavior in Online Discussion Communities", 2015

5.9.3 Future work

Although longer threads sport comparatively smaller depth, this work did not check if also *prefixes* of long threads sport comparatively smaller depths – if so, one could infer which threads will gain success.

Quotes can be seen as "higher resolution" tools than mentions, retweets and shares – in this sense it would be interesting to see if the information they provide can still be recovered in networks that only offer tools of lower resolution.

One simplifying assumption in this study is neglecting nested quotes. Although this allowed easier analysis and interpretation, removing the assumption may reveal richer information. Finally, note that this work only considers metadata associated with quotes, and does not look into the actual quote text: this leaves a wealth of possible directions open for future work.

Let me take a step back, and think of the quotes as one of the devices through which users communicate in online discussion: this chapter analyzed interaction patterns associated with quotes. From a high-level perspective, this chapter gave evidence that interaction patterns share a conventional use. In other words, interaction patterns are not only meaningless signatures, but may carry meaning beyond the individual user – e.g. may signal tie strength between

users, or status in a community. The conventions regarding the use of quotes cross the boundary of the single community: while the overall volume of quotes varies from forum to forum, how quotes are used is consistent across forums, and at various scales. I used the implicit social network of quotes between users as a main tool of analysis.

At a microscopic scale, the quote network can re-identify users in different discussions, and predict if two users in a discussion are friends with over 80% accuracy. This is remarkable if one considers that in two of the four forums friends do not quote each other directly at all. Being able to infer characteristics of the explicit social network from interaction patterns leads the way to novel applications: from automated curation of links in social networks, to socially-enhanced discussion.

At a macroscopic scale, the quote network reveals users that are prominent in the community from a real-world point of view – e.g. promoting users that were pioneers of the gaming community in RPG, and demoting bot accounts in SWZ.

Not only are interaction patterns signatures of user activity: interaction patterns also reveal the link between discussion and social structure in online communities. The next chapter concludes the analytical part of this dissertation, showing how to leverage interaction patterns to build better online environments.

6
Detecting behavior through interaction

Chapters 4 and 5 respectively showed how interaction patterns reflect personal and relational characteristics of the user. This chapter, instead, demonstrates an application of interaction patterns as indicative of user behaviour. This chapter attempts to show how interaction patterns may have an impact on some pressing problems that online communities face today.

One such major problem is abuse. Abuse is driving away core users from Twitter[1], demotivating editors on Wikipedia[2], and forcing popular sites like HackerNews[3] and Popular Science[4] to limit or disable their comment section. This chapter focuses on trolls, an extremely disruptive type of abuser[5].

Troll users operate covertly, and their abusive behaviour is non-obvious: this makes them particularly difficult to detect and contain, especially by automated systems. A growing corpus of qualitative research focuses on trolling, and differentiates it from other forms of abuse; however, its findings are not directly actionable into automated systems. On the other hand, quantitative research uses definitions of "troll" that mostly fail to capture what moderators and users consider trolling.

This work uses a different approach, relying on human moderators to obtain a gold-standard definition of troll: it takes as troll posts those that are sanctioned for trolling, and those alone. Then, it uses interaction patterns and psycholinguistic word categories to give a quantitative analysis of posts, conversations, and users sanctioned for trolling. The resulting profile of a troll user is compared to that of the civil user, and of the abuser (a user who violates the forum's rules, e.g. by flaming, but is not sanctioned for trolling). Text alone seems insufficient to detect trolls – however, a combination of lexical and interaction patterns can accurately detect trolled discussions, and reveal the responses troll generate. This analysis yields a better understanding of the behaviour of troll users, and provides useful insights for automating moderation against them.

In particular, this study makes two contributions:

- *it gives a quantitative analysis of trolling, based on a ground-truth definition by human moderators, and compares trolls to civil users and other types of abusers: this bridges the gap between the existing*

[1] Buni et al., *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*, 2016

[2] Choi et al., "Socialization tactics in wikipedia and their effects", 2010

[3] `https://techcrunch.com/2014/03/22/hacker-news-pending-comments/`, accessed on 30/1/17

[4] `http://www.popsci.com/science/article/2013-09/why-were-shutting-our-comments`, accessed on 30/1/17

[5] Enoch Peserico contributed to this work, which first appeared in Samory et al., "Sizing Up the Troll : A Quantitative Characterization of Moderator-Identified Trolling in an Online Forum", 2017

qualitative research on the topic, that calls for distinguishing trolls from other kinds of abusers, and quantitative research, that demands actionable metrics for moderating online abuse at a scale;

- *it gives evidence of why current automated moderation systems fail to detect trolls, and provides novel insights for overcoming the current limitations of such systems*

The rest of this chapter is organized as follows.

First, it details the process through which trolls are identified in the forum, and discusses the interaction patterns and text metrics used to analyse trolls' and other users' contributions.

Then it shows that, although automatically identifying trolled threads is relatively easy, accurately pinpointing trolls and trolling posts in such threads is challenging. After a comparative analysis with civil users and other abusers (both over their entire activity on the forum, and specifically when they commit infractions), the following section shows how trolls manage to remain covert while disrupting discussion: although trolls (unlike most other abusers) hardly stand out in a conversation in terms of the words they use, *how* they interact, rather than *what* they contribute, provides cues of their malicious intent. The chapter ends with a discussion of implications for moderation and future work.

6.1 Research question

Before delving into the details of the study, I remark that this work focuses on the following question:

RQ1 What linguistic metrics and interaction patterns distinguish trolls from other abusers (and civil users)?

In particular, this study makes this comparison on three levels: the corpus of posts produced by troll users in their entire activity on the forum; the specific posts that moderators sanctioned for trolling; and the discussion threads where the sanctioned posts were embedded.

6.2 What is a troll?

Moderating online content at a scale is still an open problem. While automated tools succeed in detecting barefaced forms of abuse (e.g. flaming or spamming), more sophisticated offenders elude even human moderators[6]. This is particularly true of trolls – users who create a context conducive to triggering or amplifying conflict through subtle use of aggression, deception, and/or manipulation[7]. Although there is no rigorous study on what motivates trolls, it seems that trolls create disarray for amusement's sake[8]. This is in line with recent findings, that correlate trolling and sadism[9]. Despite the prevalence and impact of trolling in computer-mediated communications [10], there is limited quantitative work that distin-

[6] Shachaf et al., "Beyond vandalism: Wikipedia trolls", 2010

[7] Hardaker, "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions", 2010; Hardaker, ""Uh. . . . not to be nitpicky,,,,but. . . the past tense of drag is dragged, not drug.": An overview of trolling strategies", 2013; Hardaker, "'I refuse to respond to this obvious troll': an overview of responses to (perceived) trolling", 2015

[8] Kirman et al., "Exploring mischief and mayhem in social computing or", 2012

[9] Buckels et al., "Trolls just want to have fun", 2014

[10] Buni et al., 2016

guishes trolling from other forms of abuse, and empirically defines how it is carried out.

6.3 Extracting trolls

This study analyses trolling in the RPG forum. Since 2012, the forum features a section devoted to public display of moderation actions: when a moderator intervenes against infractions of forum rules, a new ticket in this section reports the indicted user and post(s), along with the accusation and the disciplinary measure taken. This study categorizes posts as trolling only when the moderator explicitly phrases the accusation accordingly – i.e. when the ticket's text matches `(^|\s)troll`. Out of 1549 infringing posts recovered, 147 are trolling posts.

It is important to note that this conservative categorization of trolls may still be inaccurate[11]: moderators may misinterpret the intentions of the alleged troll, hold slightly different definitions of trolling, or fail to detect trolling altogether. However, I believe this is the most objective way to capture what the forum actually perceives as trolling behaviour.

[11] Hardaker, 2015

Hereafter this discussion refers to users as *civil*, if they do not appear in moderation tickets; *abusers*, if they appear in moderation tickets, but are never explicitly sanctioned for trolling; or *trolls*, if sanctioned at least once for trolling. Civil, abusive, and trolling posts follow the same naming convention.

6.4 Text vs. interaction metrics

This work measures text quality of user posts through metrics of *readability* (using the Automated Readability Index – ARI [12], a score that approximates the US grade level needed to comprehend a passage of text), of *politeness* (through a classifier developed in[13] for assessing civility of a request), and of *thematic coherence* (computed as the cosine similarity of the bag-of-words representation of the post with those preceding in the thread). Moreover, it analyzes post *content* matching it against the dictionaries of Linguistic Inquiry and Word Count (LIWC), a software to organize words into psychologically meaningful categories such as "inhibition" or "home"[14]. LIWC is a gold standard in psycholinguistic categorization: although its categories are quite broad and can support many different interpretations, the simultaneous over- or under-representation of *sets* of categories can often provide specific and fairly objective insights. Additionally, it analyzes non-verbal behaviour through *interaction* features – e.g. the time of posting, or the number of users in the thread – using the "content-agnostic" feature set proposed in Chapter 4 for unmasking post authors.

[12] Smith et al., "Automated readability index", 1967

[13] Danescu-Niculescu-Mizil et al., "A computational approach to politeness with application to social factors", 2013

[14] Tausczik et al., "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods", 2010

prediction task	#	interaction	text	both
troll posts	147	87 84 91	61 51 71	86 82 89
posts in trolled threads	130	93 90 95	58 48 67	92 89 94
troll posts in trolled thread	130	63 60 68	53 43 63	63 57 70
posts by trolls	1000	59 50 68	52 41 62	57 48 66

Table 6.1: Number of items in the smaller class, and percent accuracy of the four troll detection tasks using interaction features, text features, and a combination of both. Sensitivity and specificity (i.e. correct identification rate of trolls, and of others) follow accuracy values.

6.5 *Finding trolled threads is easy, trolling posts hard*

The goal of this work is to provide insight into the distinguishing features of trolls, rather than to build an accurate moderation system. But as a preliminary step, i investigate how difficult it is to detect trolls and trolling posts, and why.

I make this question concrete and quantitative by translating it into four related classification tasks. The first is the "classic" task of distinguishing trolling posts from non-trolling posts. The second is distinguishing posts *in trolled threads* from posts in non-trolled threads – note that even in trolled threads almost all posts (typically all but one) are non-trolling posts. The third task is distinguishing trolling posts from non-trolling posts in the same (trolled) thread. The fourth is distinguishing random posts made by trolls, from posts made by non-trolls – again, note that trolling posts constitute only a small minority of those posts made by trolls.

I use a combinations of textual features (LIWC counts and ARI rating), and of interaction features. I perform 10 repetitions of binary, balanced classification (i.e. for each item in the smaller class, I attempt classification with 50% probability of that item, and with 50% probability of an item chosen randomly without reinsertion from the larger class), using a Random Forest model[15], in a cross-validation scheme. The size of the smaller class and accuracy are reported in Table 6.1.

[15] Breiman, "Random forests", 2001

The classifier can identify trolling posts and trolled threads with good accuracy (respectively 87% and 93% accuracy). However, identifying trolling posts *within* a trolled thread, and "average" posts by trolls, appears considerably harder (respectively 63% and 59% accuracy). These results suggest that even when seeming to accurately detect trolling posts, the classifier is actually detecting *trolled environments*, rather than trolling posts per se – note that most non-trolling posts are in non-trolled threads. The typical behaviour of troll users does not appear significantly different from that of other users, and when trolls do act maliciously, they seem to successfully hide within a discussion that ends up uniformly "trollish".

Note that, in all tasks, textual features provide significantly less information than interaction patterns, and combining both provides little or no advantage over using interaction patterns in isolation. In other words, non-verbal behaviour may well be what can actually unmask trolls.

6.6 *Distinguishing trolls from civil users and other abusers*

The previous section showed that trolls appear hard to distinguish from other users, both in and out of trolled threads. They do have a few characteristic markers, however. This section focuses on those quantitative differences that set trolls apart from civil users, and from other abusers, *over their entire posting history*.

6.6.1 *Trolls are eager, urbane, cold-hearted contributors*

It may be surprising that on average trolls contribute to the forum over a timespan of more than 5 years, writing more than 3500 posts – significantly more than civil users, and in line with other abusers (Tukey's test, $p < .05$). Therefore, to avoid artifacts [16], I match each troll to exactly one civil user and one abuser with a similar post rate and total number of posts. This reduces the dataset to a total of roughly 1.2 million posts, authored by 120 users in each category. I then perform a series of 3-way comparisons of text and interaction features between the trolls, abusers, and civil users. All results, unless otherwise stated, are significant by Tukey's test, $p < .05$.

[16] Rosenbaum et al., "The central role of the propensity score in observational studies for causal effects", 1983

First, I focus on text quality. Trolls write less readable posts, with smaller word count and character count, compared to both abusers and civil users. This may be due to their sacrificing quality for quantity. However, posts by trolls are slightly, but not significantly, more coherent with the 3 preceding posts in the thread than those by civil users ($t = 0.607$, $p = 0.544$). Previous literature found that antisocial users tend to be less coherent than civil users [17]; this work suggest that trolls attempt to contribute useful content for a large portion of their life to gain the trust of the community [18].

[17] Cheng et al., "Antisocial Behavior in Online Discussion Communities", 2015

[18] Donath, "Idenity and Deception in the Virtual Community", 1999

Next, I examine the linguistic choices of users, measuring the frequency of LIWC categories in their posts. Abusers sport stronger use of openly offensive language than trolls and civil users, correlating negatively with "inhibition", "relative", and "social" word categories, and positively with "sexual", "death", "swear", and "bio" (body parts and biological processes) categories. Trolls, instead, just exhibit less empathy and are more confrontational, choosing fewer "inclusive", "positive affect", "future", and "tentative" words, and more "negation" and "causation" words than abusers and civil users.

Even though trolls generally talk more about personal topics (such as "money" or "work") than either civil users or abusers, they talk more than abusers but less than civil users about personal topics with stronger empathic connotations (such as "home"). This is mirrored in the different use of human-related categories. Trolls use less first and third-person pronouns than either civil users or abusers; they use more second-person and first-person-plural pronouns than abusers, but less than civil users. All this suggests trolls are eager to evoke group cohesion [19] (possibly in search for a place in the community, whether honestly or deceptively) but are

[19] Tausczik et al., 2010

less able than civil users to sustain it through empathy.

Finally, I look at the different interaction patterns of users. Trolls engage in discussion more eagerly than abusers and civil users (in terms of temporal lag from the start of conversation, number of posts preceding their first post, and propensity to write opening posts). Abusers, on the contrary, are the group with the least propensity to start conversations. Overall, trolls do not quote or get quoted differently from abusers and civil users, but they choose threads with more "intense" interaction: shorter (in terms of number of posts and time between first and last post) but more verbose (in terms of characters per post), attracting fewer views but more views per post, with participants entering the conversation earlier (in terms of number of preceding posts and inter-post lag), and with more pairs of users quoting each other.

What emerges is a profile of the troll as a user that is not *obviously* offensive, asocial or secretive, and that is in fact eager to be part of the community (indeed more than civil users) – albeit somewhat lacking in empathy towards others, and thus harsher, colder, and more confrontational.

6.6.2 *Trolls write ever more desperately*

I now focus on the changes in quality and quantity of content during the lifetime of trolls, compared to civil users and abusers. To avoid artifacts I match users as in the previous section. I then divide the activity lifespan (from first to last post) of each user into ten "ages" of equal duration, and compare readability in terms of ARI across user types and ages. Trolls and abusers enter the forum writing less readable text than civil users ($t \approx 8$, $p < 0.001$). All three user types are less readable in the last age than in the first; trolls worsen more than civil users (difference in differences via linear regression, $\beta = .191$, $p < .01$), like abusers.

Civil users see the readability of their posts improve throughout the first half of their lifetime, and slowly worsen in the second. Abusers see it worsen abruptly near the very end of their lifetime. Trolls, instead, produce posts of steadily worsening quality disseminated across an ever increasing number of threads at an ever faster pace (significantly more than civil users or abusers).

The fact that all users see the quality of their posts worsen in their last age may reflect the disaffection that eventually makes them leave the site. The sharp drop in abuser post readability may indicate a well-defined break point, that leads to a sudden departure from social norms; in fact, the majority of infractions happens around this time in an abuser's life. Existing literature confirms that readability of antisocial users starts out lower than that of other users; and it suggests that its subsequent degradation may be may be in retaliation for negative community feedback [20]. While this seems reasonable in the case of abusers, it does not fully explain why trolls would be led to post *more*, and in more threads. In fact,

[20] Cheng et al., "How community feedback shapes user behavior", 2014

it seems that the steady degradation of troll post readability is the consequence of an unexplained urge to increase their posting rate, sacrificing quality for quantity. In any case, the lack of a sharp change in posting behaviour makes trolls harder to detect than abusers.

6.7 Trolls and moderation

After examining the "normal" life of trolls, let me now focus on their actual trolling behaviour.

I begin by looking at how posts that have been moderated for trolling differ from other moderated posts, using all quality, textual, and interaction features. The language in trolling posts is more controversial (more words in the "sex", "humans" LIWC categories, $p < .05$) than that of other abusive posts. It is, however, not significantly more offensive (e.g. "swear", "negative affect" categories) or incoherent with the previous posts. Trolling posts appear earlier in the thread (in terms of wall clock time), and the conversation preceding the trolling post is more hectic (shorter timespan between posts, more users, and more posts). Overall, trolled threads receive as many replies and views as other abused threads, but in a shorter time, engaging more users, and with more user pairs exchanging quotes. The distinguishing feature of trolling posts, therefore, seems to be the level of excitement that surrounds them, rather than specific language features.

In general, trolling posts are more heavily sanctioned than other forms of abuse, considering the numeric score associated with the gravity of the penalty in the moderation tickets ($t = 2.16$, $p < .01$). However, the "criminal" history of trolls is marked by more infractions overall ($t = 4.29$ $p < .001$), and higher cumulated penalties ($t = 4.32$ $p < .001$). While few users get sanctioned for trolling as their very first post (probably intentionally created sockpuppet accounts), trolls that relapse do not troll as their first infraction. Moderators may require several rounds of sanctioning before correctly recognizing a troll[21], and despite heavier sanctions trolls remain on the site as long as other abusers after the first violation.

[21] Hardaker, 2013

6.8 Characterizing trolled threads

This section looks at trolled *threads*, giving context to troll infractions, and the reactions they provoke. All results reported are statistically significant ($p < .05$).

6.8.1 Troll posts: angst and reappraisal

I start by studying the language used by trolling posts, when compared to posts by other users in the same thread. Trolling posts are not obviously insulting (e.g. do not use more words in the "swear" LIWC category), but seem written to induce emotional responses

(more "bio", "sex", "anger", "causation", "negative" and "positive affect", "second person" words). Contrary to expectations[22], trolls do not show markers of deception – especially, trolls use complex language (more "exclusions", "prepositions", "cognitive mechanisms", as well as longer text and equal readability) that is supposedly incompatible with the cognitive load that lying requires. Finally, increased use of "causation" and "insight" is associated with reappraisal[23] - trolls may fake reconciliation (as in the case of pseudo-naive trolls[24] and "concern trolls"), or change stance in the argument[25]. In conclusion, trolling posts seem to speak to the emotionality of readers, and while they do not show signs of deception, they may mask subtle dialectic strategies. The LIWC categories associated with trolling posts sketch the troll as a hurt individual, as they find correlation in the literature with reworking of trauma, depression, and unsatisfactory relationships[26].

[22] Donath, 1999

[23] Tausczik et al., 2010

[24] Hardaker, 2013

[25] Hardaker, 2010

[26] Tausczik et al., 2010

6.8.2 *Reactions to trolls: the damage is already done*

Finally, I analyse how trolled threads evolve around the trolling posts. Posts *following* a trolling post differ from ones *preceding* it in that they feature more words in confrontational categories ("causation", "insight", "negation", "exclusive", "certainty"), and markers of debate ("past tense", and "first person singular", "second person", and "indefinite" pronouns). However, emotional charge and amount of obscene words do not differ significantly. That is to say, trolling posts (that get moderated) do not start the fire, but fan the flame.

I investigate further how the effects of trolling posts propagate across a thread. I grow a window of posts following the trolling post, and observe changes in LIWC and interaction features, compared to posts preceding the trolling post. Figure 6.1 depicts the trends in (.05 significant) t statistics for several LIWC features of interest. For posts closely following the troll post, emotional language, swear words, and sex-related words see use comparable to that in posts preceding the trolling post. However, there is a striking lessening in inhibition and inclusive language ("inclusive", "first person plural", "friends", "home"). Coincidentally, posts also become shorter and come at a slower pace, and a higher fraction of their content is quoted text. This may be an indicator of the cyclical, pointless derailments of discussion generated by trolls[27]. Widening the window of observation further from the trolling post, one can see that users return to swearing less, talking more of sensitive subjects (e.g. "money", "family", "religion"), and less of physiological processes ("body", "see"). Use of second-person pronouns increases both soon after the trolling post (possibly for accusations) and later in the thread (possibly for reappraisal). Trolling posts hide among neighbouring posts, and build upon an existing state of excitement in the discussion, to amplify controversy. Note, however, that it is possible that the "real" trolling posts, the ones that originate the

[27] Herring et al., "Searching for Safety Online: Managing "Trolling" in a Feminist Forum", 2002

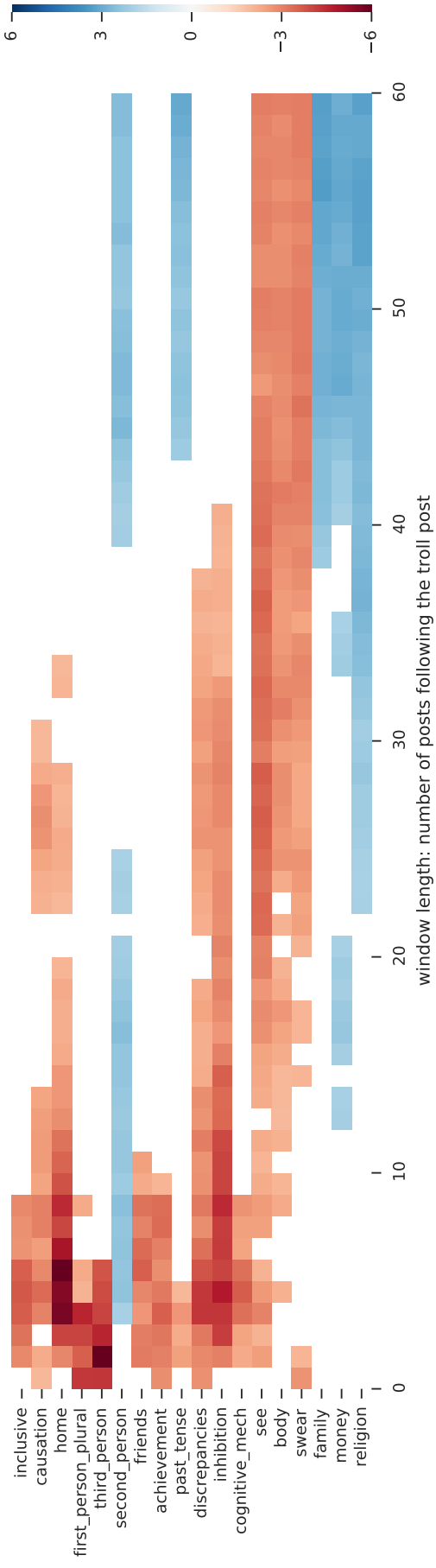


Figure 6.1: T-test statistic for selected LIWC categories, comparing posts preceding the trolling post to posts following it within a window of growing length. Colours reflects the value of the statistic, ranging from dark red (negative) to dark blue (positive). Only significant results ($p < .05$) are reported.

argument, appear earlier in the conversation yet elude moderation.

6.9 Discussion

This study quantitatively characterized moderator-identified trolling, and compared it with other forms of moderator-identified abuse and with "normal" posting activity. Although trolled environments are easy to detect, pinpointing actual culprits (both posts and users) appears much harder. First, the general conduct of trolls is less obviously uncivil than that of other abusers. Second, trolling posts, while similar to other abusive posts if analysed in isolation, are embedded in heated discussions that make them virtually indistinguishable from their context. However, interaction features can reveal discussions that will eventually be trolled, and reactions to trolling posts follow noticeable linguistic patterns.

6.10 Implications

6.10.1 Theoretical implications

This work identifies trolls via specific moderation actions, and characterizes their entire activity on the forum. This yields novel quantitative evidence that substantiates previously informal notions of troll behaviour.

First, results show that, indeed, trolls are more disruptive than other abusers: trolls receive higher sanctions for their infractions, yet they make more infractions than other abusers. Moreover, an analysis of post contents using LIWC word categories shows that trolls are less obviously offensive than other abusers. However, it is to be noted that trolls do not show particularly high scores in LIWC categories usually associated with deception.

Some results in this work support existing literature. As Hardaker[28] suggested, it seems that trolls have a history of active contribution to the forum that precedes their infractions. This may be a strategy of currying favor with the existing community to avoid being suspected when acting maliciously. Or, it may be that trolling attempts to establish an in-group of community regulars through hazing.

[28] Hardaker, 2013

Also, work by Cheng et al.[29] showed that abusers enter a community already writing less readable text than civil users, which worsens over time. Results in this work confirm this finding for both trolls and other abusers, noting that abusers' text worsens abruptly in proximity to their last posts on the forum, while trolls see their text worsen steadily throughout their lifetime on the forum.

[29] Cheng et al., 2014

Ultimately, with few notable exceptions (e.g.[30]), quantitative research on trolling largely overlooked interaction features. This work suggests they are more informative than text in detecting trolls, possibly because they are less consciously controllable than language.

[30] Mihaylov et al., "Hunting for Troll Comments in News Community Forums", 2016

6.10.2 *Practical implications*

Although detecting troll posts may appear relatively easy with "standard" techniques, what is actually easy is separating posts out of trolled threads (as all trolling posts are) from posts out of non-trolled threads (as most non-trolling posts are). Separating trolling posts from other posts *within trolled threads*, and more in general trolls from other users, is significantly harder. Results from this study suggest an alternative approach: detecting trolled *threads*, integrating longitudinal data from *user history*, and monitoring *reactions* in trolled threads to identify trolling posts. In particular, interaction features perform well in revealing discussions that will eventually be trolled, and reactions to trolling posts follow noticeable linguistic patterns. This new framing for troll detection may be directly applicable with little effort, since existing systems already have the annotated data and the tools at hand.

6.10.3 *Future work*

This work provided actionable insights on what distinguishes trolls from other users, and on how to distinguish trolling posts in the context of a discussion: it would be interesting to test these new findings by incorporating them into an automated moderation system.

Past research has often conflated generic abusers with trolls. Given that trolls are both more disruptive and less obviously uncivil than other abusive users, future research should target them as their own, separately defined category. In this regard, a large scale dataset with reliable annotations of trolling would be crucial for both theory and practice.

This work shows promising results for a quantitative characterization of trolls in *one* forum; it would be crucial to validate these findings across different platforms.

THIS CHAPTER showed the complementary strengths of interaction patterns and discussion contents in studying user behaviour. In particular, it used interaction patterns and psycholinguistic categories to give the first quantitative description of trolls, as identified by forum moderators. Interaction patterns proved to be a key asset in identifying trolls, especially when the trolls' language is misleading. This chapter showed an application of how interaction patterns can help build and maintain better environments for online discussion. The next chapter concludes this dissertation, summarizing its contributions, and suggesting how future research can benefit from them.

7
Discussion and conclusions

This dissertation started by addressing the question: *"how do we interact when we communicate through a computer?"* Its results show that the way we interact online reveals who we are, how we relate to people we interact with, and what role we play in an online community. This dissertation sheds light on the inner workings of human-computer interaction, and as such it has a potential impact on how we build and think of online discussion communities. This chapter concludes this dissertation, summarizing its contributions, suggesting possible applications, and discussing potential avenues of future research.

7.1 Contributions

The previous chapters delved into a number of specific aspects of interaction patterns. Let me now put their results into perspective, and give a bird's eye view of the main contributions of this dissertation:

1. **It provides a framework to study interaction patterns**

   Chapter 4 defines interaction patterns as content-agnostic features. This definition is both general across different forms of online interaction, and directly applicable to quantitative studies. The subsequent chapters build upon this definition, and show ways to model online activity through content-agnostic features: in particular, Chapter 4 also proposes a taxonomy for content-agnostic features, that may help future research tailor content-agnostic feature sets to specific discussion platforms. This is a fundamental step towards analysing interaction patterns as a stand-alone element of communication. In fact, while there are many interesting results in the literature that rely at least partially interaction patterns, the role of the latter has always been entangled so far with that of content, and as such difficult to assess.

2. **It proves that interaction patterns are digital signatures of the way users interact**

   Chapter 4 presents two models for authorship verification and

attribution, that respectively confirm the author of a post with 76% accuracy, and discriminate between two candidate authors with 94% accuracy. Although these models only rely on a case-study feature set of 49 hand-crafted features, the accuracy they yield is on par with that of content-based models in the literature. This has serious implications on privacy: if it is possible to identify the author of a message without reading the content of the message, many common anonymization techniques are proved ineffective. Furthermore, the feature ranking based on the weights in the authorship models is stable across four different forums: this suggests these features capture the way users discuss in forums in general, rather than in a single community. In a nutshell, interaction patterns are digital signatures of the way users interact online.

3. **It shows that interaction patterns reflect social aspects of online communities**

   Chapter 5 models the quoting activity in the forums as an implicit social network, and uses it to investigate the social structure of the forums. Not only can quotes identify users across different discussions: quotes can predict if two users in a discussion are friends with over 80% accuracy. This is remarkable, since in two of the forums friends do not share quotes with each other directly. Also, this is, to the best of my knowledge, the first attempt to link unconstrained user interactions to the explicit social network. Moreover, quotes identify prominent users in the community, better than the built-in reputation systems. This opens up opportunities to enrich discussion interfaces with social features, and to enhance social networks with alternative models of social ties. In other words, this work shows it is possible to link discussion to the social structure of online communities through interaction patterns.

4. **It shows how interaction patterns may help identify abusive behaviour**

   Chapter 6 demonstrates an application of interaction patterns to the pressing problem of moderating trolls in online communities. The customary approach to identifying trolls is learning how to distinguish troll posts from civil posts. This work suggests an alternative approach that seems more effective: finding conversations that will be trolled first, and then uncovering troll posts within these discussion. On the one hand, all posts in a trolled discussion are similar to troll posts; on the other hand, interaction patterns can detect discussions that will eventually be trolled with high accuracy, and responses to troll posts show consistent linguistic patterns. This work exemplifies how interaction patterns capture information that is distinct from – but complementary to – message content. Also, it proves how interaction patterns can reveal, beyond personal and relational aspects of users, their behaviour.

7.2 *Implications*

Having looked at the contributions of this dissertation, I would briefly discuss how it may impact current methods and theory in computer-mediated communication research.

This dissertation provides a framework to investigate online discussion through the lens of interaction patterns. Although the studies presented in Chapters 4, 5, and 6 tackle distinct research questions, results show consistent evidence that interaction patterns are informative in and of themselves, independently of the subject of each study – whether identity, relationship, or behaviour. This dissertation therefore gives a working proof that it is possible to frame problems through interaction patterns, and interaction patterns alone.

This may be crucial for generalizing the way we analyse online discussion. Since the introduction of email, text played a major role in studying online communication. However, a wealth of information flows through multimedia content (and indeed, even just through the way users interact, as this work demonstrates). In this sense, interaction patterns may be the new bag-of-words: while research is making huge leaps forward in analysing specific media (most notably spoken language and images[1]), interaction patterns provide ways to analyse online discussion independently of the medium. As an example, the taxonomy of features that Chapter 4 proposes only assumes the concepts of message and discussion, and that these may have measurable qualities in terms of time, information quantity, and interconnection.

[1] Bernardi et al., "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures", 2016

Interaction patterns may be essential metrics of user activity, when content is opaque or unavailable. However, when content is available, interaction patterns may enrich the information that content yields. Chapter 4 shows that adding language features to the authorship models trained on content-agnostic features results in performance gains (albeit relatively marginal ones). Chapter 6 shows similar results when predicting troll posts, users and threads, as well as when analysing reactions to trolls. It seems that interaction patterns carry information that is, to an extent, orthogonal to content. Interaction patterns may therefore provide an alternate viewpoint – or perhaps a magnifying glass – through which one can look at discussion.

7.3 *Applications*

This dissertation adds to the existing understanding of interaction patterns and online discussion. However, its applications are not limited to theory. This section gives few examples of how this work may help build, manage, and participate in better online discussion.

With respect to the infrastructure of discussion platforms, interaction patterns could, for example, impact current authentication procedures, exploiting the fact that interaction patterns can in fact

reveal the author of a message without looking at content. Instead of relying solely on log-in information, interaction patterns may continuously monitor that the "digital signature" of users' interactions remains coherent with that in previous sessions. A supervised model that learns users' interaction patterns may therefore serve as an unobtrusive, continuous confirmation of users' identities, as they perform ordinary activities. This could prevent online identity theft[2], or identify sockpuppet accounts used by abusive users[3]. Conversely, some platforms may want to protect their users from prying eyes by proactively altering their interaction patterns, for example introducing jitter between messages.

From a community management perspective, interaction patterns may help moderate abusive behaviour. This work gives a first quantitative profile of trolls, taking the assessment of human moderators as the gold standard for identifying them. It would be a straightforward extension of this work to incorporate this profile into an automated moderation tool that could alert human moderators of potential troublemakers. The same method and machinery could learn to distinguish other kinds of abusive behaviour besides trolling, such as spamming, flaming, or griefing.

From a user experience perspective, interaction patterns may better integrate discussion and socialization features, leveraging the fact that interaction patterns can infer if two users in a discussion are friends or not. On the one hand, interaction patterns may help curate friend lists in online social media. Friendship links may not be up to date (e.g. after a biographical break, such as relocating, or changing jobs), or may not link to friends at all[4]. Detecting friendship from interaction patterns may result in a more truthful friend list, that updates as users change the way they interact with each other. On the other hand, interaction patterns may retrofit discussion communities with social features, e.g. suggesting other users to follow, or highlighting which discussions friends are interested in.

[2] Online identity theft can have serious consequences: most notably, one tweet from a hacked account claiming explosions in the White House caused the stock market to plunge `https://www.theguardian.com/business/2013/apr/23/ap-tweet-hack-wall-street-freefall`, accessed on 30/1/17

[3] Anonymity is often considered the cause or an enabling factor of online abuse Donath, "Idenity and Deception in the Virtual Community", 1999; trolls are known for using throwaway accounts for their wrongdoings Hardaker, ""Uh. . . . not to be nitpicky,,,,but. . . the past tense of drag is dragged, not drug.": An overview of trolling strategies", 2013

[4] Ferrara et al., "The Rise of Social Bots", 2014

7.4 *Future work*

Ultimately, the goal of this dissertation is enabling future research on online discussion through the lens of interaction patterns. While I believe this dissertation sheds light on a fundamental aspect of online interaction, it also highlights unexplored areas of investigation.

One important open question is the role the communication platform plays in shaping how users interact, and in the resulting digital traces. This work focused on online forums. Future work should confirm if the findings in this study generalize to other platforms, especially platforms where content is primarily (or exclusively) non-textual – such as Pinterest, or Periscope. This would clarify when and how interaction patterns may be a useful research tool. It would also be important to study how interaction patterns change across different platforms: for example, how does the way friends

interact change when they move their conversations from one platform to another? This would probe to what extent what we observe through interaction patterns is a feature of human behaviour, rather than of the interaction platform. Furthermore, discussion often flows through several media in parallel: two people may communicate face to face, and contextually exchange pictures or directions through their phones. How does out-of-band communication affect the analytical power of interaction patterns? For example, can interaction patterns still recognize friends when half of their discussion happens offline? This would assess how robust interaction patterns are in a real-world application scenario. Concurrently, research should also account for the bias that interface imparts on interaction: for instance, if only few friends are directly visible in a user's friend list, it is more likely he will interact with them. De-biasing interaction patterns from the effects of presentation would be useful for most social computing research.

Future research should also consider different methods for analysing interaction patterns. This dissertation measured them through feature engineering; extracting features from scratch, in an unsupervised manner, may uncover non-obvious interaction patterns[5]. Moreover, this work measured all interaction at the level of posts and quotes, since these were the minimal unit of information exchange in a discussion for the purposes of this study, and analysed interaction patterns as a collection of these units. Nonetheless, more complex models of interaction patterns could uncover finer details of user activity. For example, a hierarchical model could better deal with information with different levels of granularity, such as post-, thread-, and user-level features; a longitudinal model could better investigate how a discussion evolves, or how users gain status.

[5] Zhang et al., "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification", 2015

An open question is how to evaluate multimodal interaction. Interaction patterns may be measured through features that differ in nature; for example, timing or topology. How can we measure when text is more informative than timing? And by how much? This work addresses such questions by contrasting the weights for different feature sets as learned by a model, or selectively adding feature sets and comparing the prediction accuracy they yield. However, this approach has limitations, since it only assesses how well the model is able to exploit the features, and not the relative performance of different features in general.

Finally, the main contribution of this work is foundational: it proves that interaction patterns carry information in and of themselves, and it showcases how they can be used to analyse higher level constructs, such as status and friendship. However, there are vast opportunities of future research for targeting other constructs. How does the personality of users affect the way they interact? How do interaction patterns reflect social support? What interaction patterns are signals of social identity?

A
Appendix: quote network features

Most features for the tasks of user re-identification and friendship prediction were extracted using – and are described in the documentation of – the python module networkx[1] .

[1] http://networkx.readthedocs.io/en/networkx-1.11/reference/algorithms.html

A.1 Quote network metrics for user fingerprinting

degree
in degree
out degree
self loops
number of triangles
clustering coefficient
square clustering coefficient
assortative mixing (all combinations of in and out degrees)
average neighbor degree
number of edges in ego network
number of nodes in ego network
ego network density
HITS: hubs, authorities
PageRank
transitivity
eccentricity
vitality
closeness vitality
betweenness centrality
degree centrality
closeness centrality
Katz centrality
communicability centrality
load centrality
eigenvector centrality
current flow betweenness centrality
current flow closeness centrality

A.2 Quote network metrics for friend prediction

Note that in the description of the features I use the term "friend" as a shorthand for users linked through quotes: it should not be confused with "friend" as in the forum friendship system.

number of directed edges in the pair
number of common friends
average clustering of common friends
number of edges between common friends
reciprocal of the fraction of edges that are not reciprocated
reciprocity weighted by the out-degree of the nodes
ratio of the minimum and the maximum of the edges in one direction among the pair
fraction of the edges of the two nodes that are within the pair
assortative mixing of the common friends
minimum and maximum of the dispersion of the nodes in the pair
minimum and maximum number of edges in one direction within the pair
minimum and maximum of the average neighbor degrees for the nodes in the pair
Jaccard coefficient
preferential attachment
resource allocation index
Adamic Adar index
number of common threads
Jaccard index of the common threads
delta measure on the number of authors in the common threads
Adamic Adar index on the number of authors in the common threads
sum of reciprocals of the number of authors in the common threads
product of the number of threads for both nodes in the pair

Bibliography

Abbasi, Ahmed and Hsinchun Chen. "Applying authorship analysis to extremist-group web forum messages". In: *Intelligent Systems, IEEE* 20.5 (2005), pp. 67–75.

– "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace". In: *ACM Transactions on Information Systems* 26.2 (2008).

Abdallah, Emad E, A.F. Otoom, Ola Abu-aisheh, Diana Omari, and Ghadeer Salem. "Detecting Email Forgery using Random Forests and Naïve Bayes Classifiers". In: *International Science Index* 6.3 (2012), pp. 276–280.

Achlioptas, Dimitris, Aaron Clauset, David Kempe, and Cristopher Moore. "On the bias of traceroute sampling". In: *Journal of the ACM* 56.4 (June 2009), pp. 1–28.

Aggarwal, Charu and Karthik Subbian. "Evolutionary network analysis: A survey". In: *ACM Computing Surveys (CSUR)* 47.1 (2014), pp. 1–36.

Aiello, Luca Maria, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. "Friendship prediction and homophily in social media". In: *ACM Transactions on the Web* 6.2 (2012), pp. 1–33.

Akker, Rieks op den and David Traum. "A comparison of addressee detection methods for multiparty conversations". In: *Diaholmia 2009 (semdial 2009)*. 2009.

Alis, Christian M and May T Lim. "Spatio-temporal variation of conversational utterances on Twitter." In: *PloS one* 8.10 (2013), e77793.

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. "Discovering value from community activity on focused question answering sites: a case study of stack overflow". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'12*. 2012.

Anwar, Tarique and Muhammad Abulaish. "Modeling a web forum ecosystem into an enriched social graph". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8329 LNAI (2013), pp. 152–172.

Arakawa, Yui, Akihiro Kameda, Akiko Aizawa, and Takafumi Suzuki. "Adding twitter-specific features to stylistic features for classifying tweets by user type and number of retweets". In:

Journal of the Association for Information Science and Technology 65.7 (2014), pp. 1416–1423.

Aumayr, Erik, Jeffrey Chan, and Conor Hayes. "Reconstruction of Threaded Conversations in Online Discussion Forums". In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. 2011, pp. 26–33.

Backstrom, Lars, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. "Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry". In: *Proceedings of the 6th ACM International Conference on Web Search and Data Mining - WSDM '13*. 2013, pp. 13–22.

Backstrom, Lars and Jon Kleinberg. "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2014), pp. 831–841.

Barabási, Albert-László. "The origin of bursts and heavy tails in human dynamics". In: *Nature* 435.May (2005).

Barcellini, Flore, Françoise Détienne, Jean Marie Burkhardt, and Warren Sack. "A socio-cognitive analysis of online design discussions in an Open Source Software community". In: *Interacting with Computers* 20.1 (2008), pp. 141–165.

– "A study of online discussions in an open-source software: Community reconstructing thematic coherence and argumentation from quotation practices". In: *Proceedings of the 2nd Communities and Technologies Conference, C and T 2005* July (2005), pp. 301–320.

Bernardi, Raffaella, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures". In: *Journal of Artificial Intelligence Research* 55.1 (2016), pp. 1–34.

Bogdanova, Dasha and Angeliki Lazaridou. "Cross-Language Authorship Attribution". In: *Proceedings of the International Conference on Language Resources and Evaluation*. 2014, pp. 2015–2020.

boyd, danah, Scott Golder, and Gilad Lotan. "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter". In: *Proceedings of the Annual Hawaii International Conference on System Sciences*. 2010, pp. 1–10.

Breiman, Leo. "Random forests". In: *Machine learning* (2001), pp. 5–32.

Brin, Sergey and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine". In: *Computer Networks and ISDN Systems* 30.1-7 (Apr. 1998), pp. 107–117.

Brocardo, Marcelo Luiz, Issa Traore, and Isaac Woungang. "Authorship verification of e-mail and tweet messages applied for continuous authentication". In: *Journal of Computer and System Sciences* 1 (Dec. 2014), pp. 1–12.

Buckels, Erin E., Paul D. Trapnell, and Delroy L. Paulhus. "Trolls just want to have fun". In: *Personality and Individual Differences* 67 (Sept. 2014), pp. 97–102.

Buni, Catherine and Soraya Chemaly. *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*. 2016.

Burke, Moira, Cameron Marlow, and Thomas Lento. "Social Network Activity and Social Well-Being". In: *Proceedings of the International Conference on Human Factors in Computing Systems - CHI '10* (2010), pp. 1909–1912.

Burrows, John. "'Delta': a measure of stylistic difference and a guide to likely authorship". In: *Literary and Linguistic Computing* 17.3 (2002), pp. 267–287.

Canseco, Leonardo, Lori Lamel, and Jean-Luc Gauvain. "A comparative study using manual and automatic transcriptions for diarization". In: *IEEE Workshop on Automatic Speech Recognition and Understanding* (2005), pp. 415–419.

Chandrasekharan, Eshwar, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. "The Bag of Communities Approach : Identifying Abusive Behavior Online with Preexisting Internet Data". In: *ACM CHI Conference on Human Factors in Computing Systems*. 2017.

Chang, Shuo, Vikas Kumar, Eric Gilbert, and Loren G. Terveen. "Specialization, Homophily, and Gender in a Social Curation Site: Findings from Pinterest". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '14*. New York, New York, USA: ACM Press, 2014, pp. 674–686.

Chaudhuri, Sourish and Bhiksha Raj. "A comparison of latent variable models for conversation analysis". In: *Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (2011), pp. 30–38.

Cheng, Justin, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "Antisocial Behavior in Online Discussion Communities". In: *Proceedings of the Ninth International AAAI Conference on Web and Social Media - ICWSM '15*. Apr. 2015.

– "How community feedback shapes user behavior". In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media - ICWSM '14*. 2014, pp. 41–50.

Choi, Boreum, Kira Alexander, Robert E Kraut, and John M Levine. "Socialization tactics in wikipedia and their effects". In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*. March. New York, New York, USA: ACM Press, 2010, p. 107.

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. "A computational approach to politeness with application to social factors". In: *Proceedings of ACL* (2013).

De Choudhury, Munmun, Winter a. Mason, Jake M. Hofman, and Duncan J. Watts. "Inferring relevant social networks from interpersonal communication". In: *Proceedings of the 19th International Conference on World Wide Web - WWW '10* (2010), p. 10.

De Vel, Olivier. "Mining e-mail authorship". In: *Information Retrieval* 30.4 (2000), p. 55.

De Vel, Olivier, Alison Anderson, Malcolm Walter Corney, and George Mohay. "Mining e-mail content for author identification forensics". In: *ACM SIGMOD Record*. Vol. 30. 4. 2001, p. 55.

Dino, A., S. Reysen, and Nyla R Branscombe. *Online Interactions Between Group Members Who Differ in Status*. 2008.

Donath, Judith. "Idenity and Deception in the Virtual Community". In: *Communities in Cyberspace* (1999), pp. 27–58.

– "Signals , cues and meaning". 2011.

Eder, M. "Does size matter? Authorship attribution, small samples, big problem". In: *Digital Scholarship in the Humanities* (2014).

Epstein, Robert and Ronald E Robertson. "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.33 (2015), pp. 4512–21.

Farajtabar, Mehrdad, Yichen Wang, Manuel Gomez-Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. "COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution". In: *in Proceedings of the 28th International Conference on Neural Information Processing Systems - NIPS '15*. Ed. by C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett. Curran Associates, Inc., 2015, pp. 1945–1953.

Ferrara, Emilio. "A large-scale community structure analysis in Facebook". In: *EPJ Data Science* 1.1 (Dec. 2012), p. 9.

Ferrara, Emilio and Zeyao Yang. "Measuring emotional contagion in social media". In: *PLoS ONE* 10.11 (2015), pp. 1–14.

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. "The Rise of Social Bots". In: *arXiv preprint arXiv:1407.5225* grant 220020274 (2014), pp. 1–11.

Frey, Davide, Arnaud Jégou, and Anne Marie Kermarrec. "Social market: Combining explicit and implicit social networks". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6976 LNCS (2011), pp. 193–207.

Friggeri, Adrien, LA Adamic, Dean Eckles, and Justin Cheng. "Rumor Cascades". In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media - ICWSM '14*. 2014, pp. 101–110.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. "Quantifying Controversy in Social Media". In: *arXiv preprint arXiv:1507.05224* (2015).

Garimella, Kiran, Ingmar Weber, and Munmun De Choudhury. "Quote RTs on Twitter". In: *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*. 2016, pp. 200–204.

Gatica-Perez, Daniel. "Automatic nonverbal analysis of social interaction in small groups: A review". In: *Image and Vision Computing* 27.12 (Nov. 2009), pp. 1775–1787.

Gilbert, Eric. "Predicting tie strength in a new medium". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12* (2012), p. 1047.

Gilbert, Eric and Karrie Karahalios. "Predicting tie strength with social media". In: *ACM Conference on Human Factors in Computing Systems*. 2009, pp. 211–220.

Gómez, Vicenç, Andreas Kaltenbrunner, and Vicente López. "Statistical analysis of the social network and discussion threads in slashdot". In: *Proceeding of the 17th international conference on World Wide Web - WWW '08* (2008), p. 645.

Govindan, Priya, Jin Xu, Shawndra Hill, Tina Eliassi-Rad, and Chris Volinsky. "Local Structural Features Threaten Privacy across Social Networks". In: *The 5th Workshop on Information in Networks*. 2013.

Graells-Garrido, Eduardo, Mounia Lalmas, and Ricardo Baeza-Yates. "Data Portraits and Intermediary Topics: Encouraging Exploration of Politically Diverse Profiles". In: *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI '16* Iui (2016), pp. 228–240.

Granovetter, Mark S. "The strength of weak ties: A network theory revisited". In: *Sociological theory* 1.1983 (1983), pp. 201–233.

Gupte, Mangesh and Tina Eliassi-Rad. "Measuring tie strength in implicit social networks". In: *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12* (2012), pp. 109–118.

Halevy, Alon, Peter Norvig, and Fernando Pereira. "The Unreasonable Effectiveness of Data". In: *IEEE Intelligent Systems* 24.2 (2009), pp. 8–12.

Han, Kyungsik, Sanghack Lee, Jin Yea Jang, Yong Jung, and Dongwon Lee. "" Teens are from Mars , Adults are f rom Venus " : Analyzing and Predicting Age Groups with Behavioral Characteristics in Instagram". In: *ACM Web Science 2016* (2016), pp. 35–44.

Hardaker, Claire. "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions". In: *Journal of Politeness Research* 6.2 (2010), pp. 215–242.

– "'I refuse to respond to this obvious troll': an overview of responses to (perceived) trolling". In: *Corpora* 10.2 (Aug. 2015), pp. 201–229.

– ""Uh. . . . not to be nitpicky,,,,,but. . . the past tense of drag is dragged, not drug.": An overview of trolling strategies". In: *Journal of Language Aggression and Conflict* 1.1 (2013), pp. 58–86.

Herring, Susan, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. "Searching for Safety Online: Managing "Trolling" in a Feminist Forum". In: *The Information Society* 18.5 (2002), pp. 371–384.

Hum, Noelle J., Perrin E. Chamberlin, Brittany L. Hambright, Anne C. Portwood, Amanda C. Schat, and Jennifer L. Bevan. "A picture is worth a thousand words: A content analysis of Facebook profile photographs". In: *Computers in Human Behavior* 27.5 (2011), pp. 1828–1833.

Hutto, C. J., Sarita Yardi, and Eric Gilbert. "A longitudinal study of follow predictors on twitter". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (2013), p. 821.

Juola, Patrick. "Authorship Attribution". In: *Foundations and Trends® in Information Retrieval* 1.3 (2007), pp. 233–334.

– "Future trends in authorship attribution". In: *Advances in digital forensics III* 242 (2007), pp. 119–132.

Kang, Jeon Hyung and Jihie Kim. "Analyzing answers in threaded discussions using a role-based information network". In: *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011* (2011), pp. 111–117.

Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network". In: *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '03*. 2003, pp. 137–146.

Kirman, Ben, Conor Lineham, and Shaun Lawson. "Exploring mischief and mayhem in social computing or". In: *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts - CHI EA '12*. New York, New York, USA: ACM Press, 2012, p. 121.

Koessler Gosnell, Denise. "Social Fingerprinting : Identifying Users of Social Networks by their Data Footprint". PhD thesis. The University of Tennessee, Knoxville, 2014.

Koppel, Moshe. "Automatically Categorizing Written Texts by Author Gender". In: *Literary and Linguistic Computing* 17.4 (2002), pp. 401–412.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. "Authorship Attribution: What's Easy and What's Hard?" In: *Journal of Law and Policy* 39.2006 (2013), pp. 317–331.

– "Computational Methods in Authorship Attribution". In: *International Review of Research in Open and Distance Learning* 14.4 (2008), pp. 90–103.

Koppel, Moshe, Jonathan Schler, and Elisheva Bonchek-Dokow. "Measuring differentiability: unmasking pseudonymous authors". In: *Journal of Machine Learning Research* 8 (2007), pp. 1261–1276.

Koppel, Moshe, Jonathan Schler, Shlomo Argamon, and Yaron Winter. "The "Fundamental Problem" of Authorship Attribution". In: *English Studies* 93.3 (2012), pp. 284–291.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massivescale emotional contagion

through social networks". In: *Proceedings of the National Academy of Sciences* 111.29 (July 2014), pp. 10779–10779.

Kumar, Ravi, Mohammad Mahdian, and Mary McGlohon. "Dynamics of Conversations". In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10*. 2010, pp. 553–562.

Kunii, Y. and H. Hashimoto. "Tele-handshake using HandShake Device". In: *Proceedings of IECON '95 - 21st Annual Conference on IEEE Industrial Electronics* 1 (1995).

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?" In: *Proceedings of the 19th international conference on World wide web - WWW '10*. 2010, p. 591.

Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. "Graphs over time: Densification Laws, Shrinking Diameters and Possible Explanations". In: *Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD '05*. New York, New York, USA: ACM Press, 2005, p. 177.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. "Meme-tracking and the Dynamics of the News Cycle". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '09*. 2009, 497–506.

Leskovec, Jure, Anand Rajaraman, and Jeffrey D. Ullman. *Mining of massive datasets*. 2014.

Leskovec, Jure and Eric Horvitz. "Planetary-scale views on a large instant-messaging network". In: *Proceeding of the 17th international conference on World Wide Web - WWW '08* (2008), p. 915.

Li, Jiexun, Rong Zheng, and Hsinchun Chen. "From fingerprint to writeprint". In: *Communications of the ACM* 49.4 (2006), pp. 76–82.

Li, Weifeng, Ahmed Abbasi, Shiyu Hu, and Victor Benjamin. "Modeling Interactions in Web Forums". In: *Proceedings of the 2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference*. 2014, pp. 1–9.

Liben-Nowell, David and Jon Kleinberg. "The link-prediction problem for social networks". In: *Journal of the American Society for Information Science and Technology* 58.7 (May 2007), pp. 1019–1031.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

Meusel, Robert, Sebastiano Vigna, Oliver Lehmberg, and Christian Bizer. "Graph structure in the web — revisited". In: *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*. New York, New York, USA: ACM Press, 2014, pp. 427–432.

Mihaylov, Todor and Preslav Nakov. "Hunting for Troll Comments in News Community Forums". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2016), pp. 399–405.

Mislove, Alan, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. "Measurement and Analysis of Online Social Networks". In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement - IMC '07* (2007), pp. 29–42.

Mitra, Tanushree and Eric Gilbert. "Analyzing Gossip in Workplace Email". In: *ACM SIGWEB Newsletter* Winter (Jan. 2013), pp. 1–7.

Mohtasseb, Haytham and Amr Ahmed. "More blogging features for author identification". In: *The 2009 International Conference on Computer Engineering and Applications*. 2009, pp. 461–466.

Montjoye, Yves-Alexandre de, César a Hidalgo, Michel Verleysen, and Vincent D Blondel. "Unique in the Crowd: The privacy bounds of human mobility." In: *Scientific reports* 3 (2013), p. 1376.

Myers, Seth A, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. "Information network or social network? The Structure of the Twitter Follow Graph". In: *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*. New York, New York, USA: ACM Press, 2014, pp. 493–498.

Narayanan, Arvind and Vitaly Shmatikov. "De-anonymizing social networks". In: *Proceedings of the 2009 IEEE Symposium on Security and Privacy*. Ieee, May 2009, pp. 173–187.

Narayanan, Arvind, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. "On the Feasibility of Internet-Scale Author Identification". In: *Proceedings of the 2012 IEEE Symposium on Security and Privacy* (2012), pp. 300–314.

Narayanan, Arvind and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy* (May 2008), pp. 111–125.

Newman, M. E. J. and Juyong Park. "Why social networks are different from other types of networks". In: *Physical Review E* 68.3 (Sept. 2003), p. 036122.

Niculae, Vlad, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns". In: *Proceedings of the 24th International Conference on World Wide Web - WWW '15*. 2015, pp. 798–808.

Noecker, John, Michael Ryan, and Patrick Juola. "Psychological profiling through textual analysis". In: *Literary and Linguistic Computing* 28.3 (2013), pp. 382–387.

Owens, David, Robert Sutton, and Margaret a Neale. "Technologies of Status Negotiation: Status Dynamics in Email Discussion Groups". In: *Research in Managing Groups and Teams* 3 (2000).

Panzarasa, Pietro, Tore Opsahl, and Kathleen M. Carley. "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community". In: *Journal of the American Society for Information Science and Technology* 60.5 (2009), pp. 911–932.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel,

Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2012), pp. 2825–2830.

Peng, Fuchun, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. "Language independent authorship attribution using character level language models". In: *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1* (2003), 267–274.

Pillay, Sangita R. and Thamar Solorio. "Authorship attribution of web forum posts". In: *General Members Meeting and eCrime Researchers Summit, eCrime 2010*. 2010.

Pratanwanich, Naruemon and Pietro Lio'. "Who Wrote This? Textual Modeling with Authorship Attribution in Big Data". In: *The 2nd International Workshop on High Dimensional Data Mining*. 2014.

Rheingold, Howard. *The Virtual Community*. 1986.

Rosenbaum, Paul R. and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

Roth, Maayan, Guy Flysher, Yossi Matias, Ari Leichtberg, and Ron Merom. "Suggesting Friends Using the Implicit Social Graph". In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10*. 2010.

Rudman, Joseph. "The State of Non-Traditional Authorship Attribution Studies—2012: Some Problems and Solutions". In: *English Studies* 93.3 (2012), pp. 259–274.

Samory, Mattia, Federica Bogo, and Enoch Peserico. "Community structure and interaction dynamics through the lens of quotes". In: *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*. 2016, pp. 358–359.

Samory, Mattia and Enoch Peserico. "Content attribution ignoring content". In: *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*. 2016, pp. 233–243.

– "Quotes in forum.rpg.net". In: *Proceedings of the ACM Conference on Web Science - WebSci '15*. 2015, pp. 1–2.

Samory, Mattia, Vincenzo-Maria Cappelleri, and Enoch Peserico. "Quotes Reveal Community Structure and Interaction Dynamics". In: *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing - CSCW' 17*. 2017.

Samory, Mattia and Enoch Peserico. "Sizing Up the Troll : A Quantitative Characterization of Moderator-Identified Trolling in an Online Forum". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '17*. 2017.

Seroussi, Yanir, Ingrid Zukerman, and Fabian Bohnert. "Authorship Attribution with Latent Dirichlet Allocation". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. 2011, pp. 181–189.

Shachaf, Pnina and Noriko Hara. "Beyond vandalism: Wikipedia trolls". In: *Journal of Information Science* 36.3 (2010), pp. 357–370.

Shafiq, M. Zubair, Muhammad U. Ilyas, Alex X. Liu, and Hayder Radha. "Identifying leaders and followers in online social networks". In: *IEEE Journal on Selected Areas in Communications* 31.9 (2013), pp. 618–628.

Shriberg, Elizabeth. "Spontaneous speech: How people really talk and why engineers should care". In: *Proceedings of the 9th European Conference on Speech Communication and Technology - Interspeech 2005*. 2005, pp. 1781–1784.

Smith, Edgar A. and R. J. Senter. "Automated readability index". In: *AMRL-TR. Aerospace Medical Research Laboratories (U.S.)* (May 1967), pp. 1–14.

Soni, Sandeep, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. "Modeling Factuality Judgments in Social Media Text". In: *ACL*. 2014, pp. 415–420.

Stamatatos, Efstathios. "A survey of modern authorship attribution methods". In: *Journal of the American Society for Information Science and Technology* 60.3 (2009), pp. 538–556.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. "Computer-based Authorship Attribution without Lexical Measures". In: *Computers and the Humanities* (2001), pp. 193–214.

Sumi, Yasuyuki, Sadanori Ito, Tetsuya Matsuguchi, Sidney Fels, Shoichiro Iwasawa, Kenji Mase, Kiyoshi Kogure, and Norihiro Hagita. "Collaborative capturing, interpreting, and sharing of experiences". In: *Personal and Ubiquitous Computing* 11.4 (Sept. 2006), pp. 265–271.

Tausczik, Yla R. and James W. Pennebaker. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods". In: *Journal of Language and Social Psychology* 29.1 (2010), pp. 24–54.

Tufekci, Zeynep. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls". In: *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media - ICWSM '14* (2014), p. 10.

Welser, Howard T, Eric Gleave, Danyel Fisher, and Mark Smith. "Visualizing the signatures of social roles in online discussion groups". In: *Journal of social structure* 8.2 (2007), pp. 1–32.

Wilson, Christo, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. "User interactions in social networks and their implications". In: *Proceedings of the fourth ACM european conference on Computer systems - EuroSys '09* (2009), p. 205.

Wilson, R. E., S. D. Gosling, and L. T. Graham. "A Review of Facebook Research in the Social Sciences". In: *Perspectives on Psychological Science* 7.3 (2012), pp. 203–220.

Zhang, Chunxia, Xindong Wu, Zhendong Niu, and Wei Ding. "Authorship identification from unstructured texts". In: *Knowledge-Based Systems* 66 (2014), pp. 99–111.

Zhang, Ye and Byron Wallace. "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification". In: 1 (2015).

Zheng, Rong, Jiexun Li, Hsinchun Chen, and Zan Huang. "A framework for authorship identification of online messages: Writing-style features and classification techniques". In: *Journal of the American Society for Information Science and Technology* 57.3 (2006), pp. 378–393.

Zhou, Wei, Wenjing Duan, and Selwyn Piramuthu. "A social network matrix for implicit and explicit social network plates". In: *Decision Support Systems* 68 (2014), pp. 89–97.