Community Identity and User Engagement in a Multi-Community Landscape

Community Identity and User Engagement in a Multi-Community Landscape

Justine Zhang*
Cornell University
jz727@cornell.edu

William L. Hamilton*
Stanford University
wleif@stanford.edu

Cristian Danescu-Niculescu-Mizil
Cornell University
cristian@cs.cornell.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Abstract

A community's identity defines and shapes its internal dynamics. Our current understanding of this interplay is mostly limited to glimpses gathered from isolated studies of individual communities. In this work we provide a systematic exploration of the nature of this relation across a wide variety of online communities. To this end we introduce a quantitative, language-based typology reflecting two key aspects of a community's identity: how *distinctive*, and how temporally *dynamic* it is. By mapping almost 300 Reddit communities into the landscape induced by this typology, we reveal regularities in how patterns of user engagement vary with the characteristics of a community.

Our results suggest that the way new and existing users engage with a community depends strongly and systematically on the nature of the collective identity it fosters, in ways that are highly consequential to community maintainers. For example, communities with distinctive and highly dynamic identities are more likely to retain their users. However, such niche communities also exhibit much larger acculturation gaps between existing users and newcomers, which potentially hinder the integration of the latter.

More generally, our methodology reveals differences in how various social phenomena manifest across communities, and shows that structuring the multi-community landscape can lead to a better understanding of the systematic nature of this diversity.

1 Introduction

"If each city is like a game of chess, the day when I have learned the rules, I shall finally possess my empire, even if I shall never succeed in knowing all the cities it contains."
— Italo Calvino, Invisible Cities

A community's identity—defined through the common interests and shared experiences of its users—shapes various facets of the social dynamics within it (Ren, Kraut, and Kiesler 2007; Tajfel 2010; Ren et al. 2012). Numerous instances of this interplay between a community's identity and social dynamics have been extensively studied in the context of individual online communities (Bryant, Forte, and Bruckman 2005; Lampe et al. 2010; Danescu-Niculescu-Mizil et al. 2013). However, the sheer variety of online platforms complicates the task of generalizing insights beyond these isolated, single-community glimpses. A new way to reason about the variation across multiple communities is needed in order to systematically characterize the relationship between properties of a community and the dynamics taking place within.

One especially important component of community dynamics is user engagement. We can aim to understand why users join certain communities (Panciera, Halfaker, and Terveen 2009), what factors influence user retention (Dror et al. 2012), and how users react to innovation (Danescu-Niculescu-Mizil et al. 2013). While striking patterns of user engagement have been uncovered in prior case studies of individual communities (Postmes, Spears, and Lea 2000; Huffaker et al. 2006; Fugelstad et al. 2012; Otterbacher and Hemphill 2012; McAuley and Leskovec 2013), we do not know whether these observations hold beyond these cases, or when we can draw analogies between different communities. Are there certain types of communities where we can expect similar or contrasting engagement patterns?

To address such questions quantitatively we need to provide structure to the diverse and complex space of online communities. Organizing the multi-community landscape would allow us to both characterize individual points within this space, and reason about systematic variations in patterns of user engagement across the space.

Present work: Structuring the multi-community space. In order to systematically understand the relationship between community identity[1] and user engagement we introduce a quantitative typology of online communities. Our typology is based on two key aspects of community identity: how *distinctive*—or niche—a community's interests are relative to other communities, and how *dynamic*—or volatile—these interests are over time. These axes aim to capture the salience of a community's identity and dynamics of its temporal evolution.

*The two first authors contributed equally and are ordered non-alphabetically to balance ordering in another collaboration.
Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]We use "community identity" and "collective identity" interchangeably to refer to the shared definition of a group, derived from members' common interests and shared experiences. We are not directly concerned with the more sociopolitical and psychological connotations of these terms (Polletta and Jasper 2001; Simon and Klandermans 2001; Ashmore, Deaux, and McLaughlin-Volpe 2004).

Justine Zhang
Cornell University

William L. Hamilton
Stanford University

Cristian Danescu-Niculescu-Mizil
Cornell University

Dan Jurafsky
Stanford University

Jure Leskovec
Stanford University

ICWSM 2017

Online Community Identity

- A **quantitative, language based** typology, Reflecting two aspects:
  - Distinctiveness
  - Dynamicity

- How different types vary in **User Engagement**:
  - **User retention** (proportion of users who contribute at month t and then in t+1)
  - attract **New Members**

Methodology

- Word based metrics for **Specificity** and **Volatility**
  - based on comparison of **word Freq**. in a setting vs. some background distribution
    - **P**ointwise **M**utual **I**nformation (**PMI**)

$$\operatorname{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}.$$

probability of their coincidence given their joint distribution and their individual distributions

Specificity (SC=0 : generic)

$$\mathcal{S}_c(w) = \log \frac{\mathrm{P}_c(w)}{\mathrm{P}_C(w)},$$

c: Community
C : all Communities

Volatility (V ~0 : stable)

$$\mathcal{V}_{c_t}(w) = \log \frac{\mathrm{P}_{c_t}(w)}{\mathrm{P}_{c_T}(w)}.$$

t: single Snapshot
T : all Snapshots

Methodology

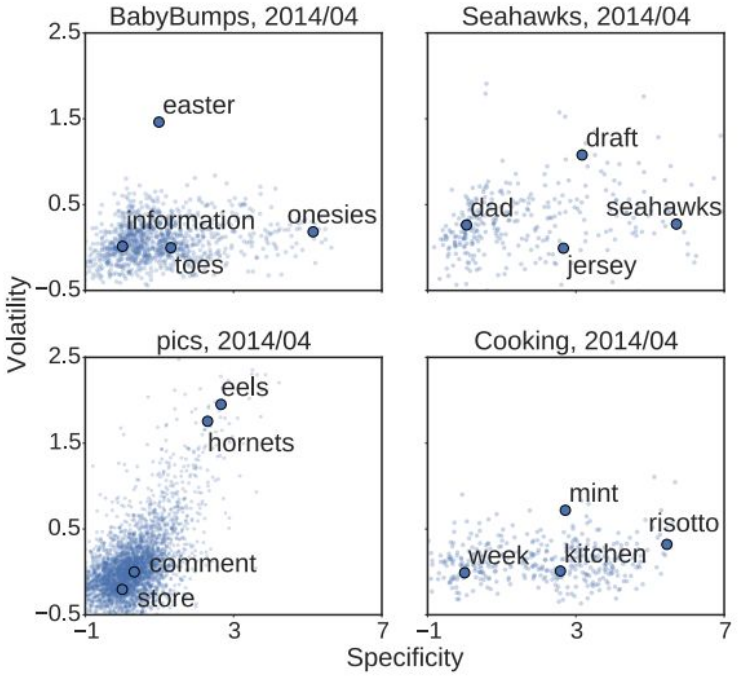
- From Words to Community-Level Measures
  - Extending to utterances:
    - Sc of utterance d would be the **Average** of all words in d (same for V)
  - Distinctiveness: **N(c)**
    - Average of S for all utterances in **c** in time **t**
  - Dynamicity: **D(c)**
    - Average of V for all utterances in **c** at time **t**

Reddit Communities : Some Examples

- **Seahawks**
  - Sport fans
- **BabyBumps**
  - expecting Mothers
- **Cooking**
  - recipe and ideas
- **Pics**
  - image sharing discussion

Reddit Communities : Some Examples

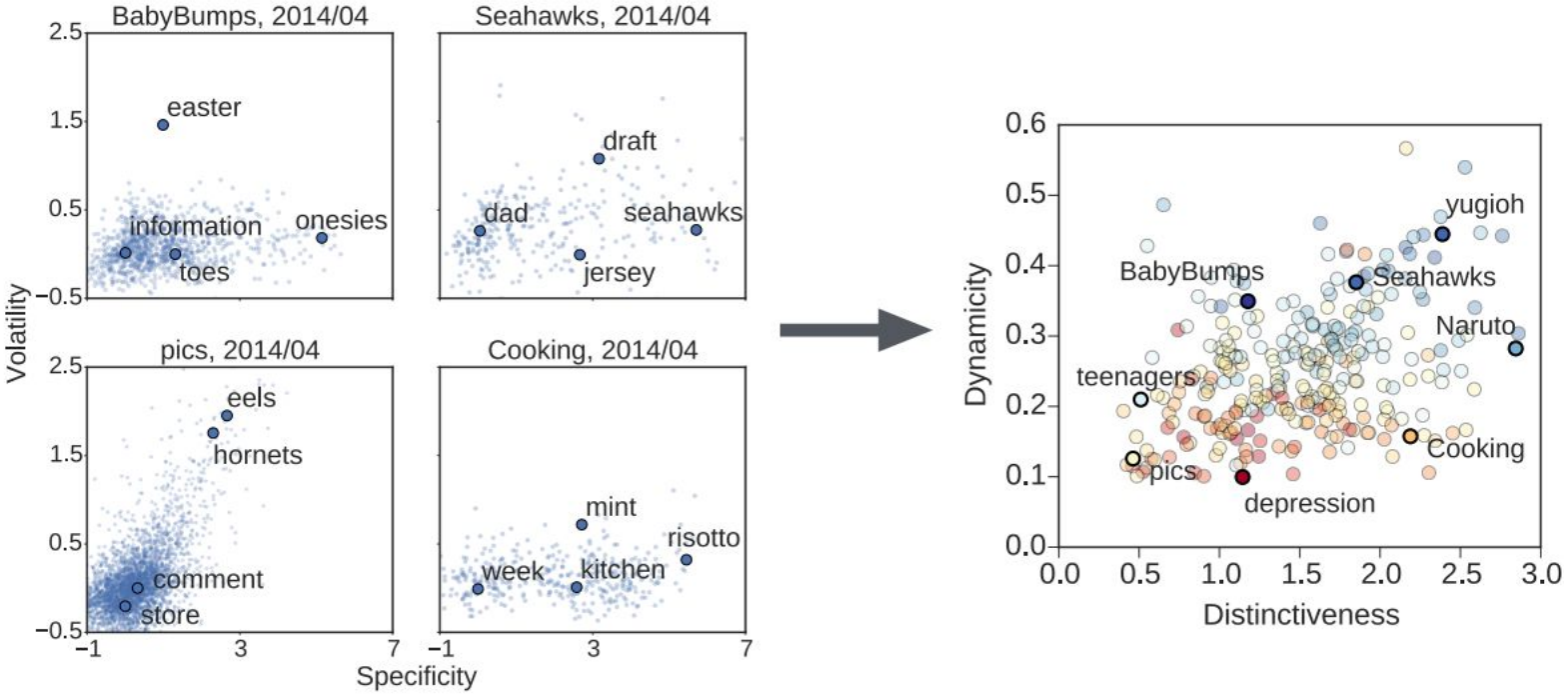
- **BabyBumps**
  - Not consistent but dominant for a while (easter)
- **Cooking**
  - Specific topics: e.g., *Risotto*
  - Dynamic:
    kitchens vs. *Mint*
    (constant topic) vs. (seasonal topic)



DataSet

- All Subreddits from **Jan 2013** to **December 2014** (user activities up to may 2015)
  - All with at least **500** words in at least **4 months** of subReddits's history
  - Time - window : Monthly
  - Communities in foreign languages?
    - Manually removed
  - Overall : **283 Community**
    - Considering only Top-level Comments (initial responses to a post)
    - keeping set of (word,user) so so word level data is not skewed by highly active users
    - Focusing only on Nouns
    - Discarding long tail of infrequent words (focusing on the top 5 percentile of words)

Methodology



Reddit Communities : Some Examples

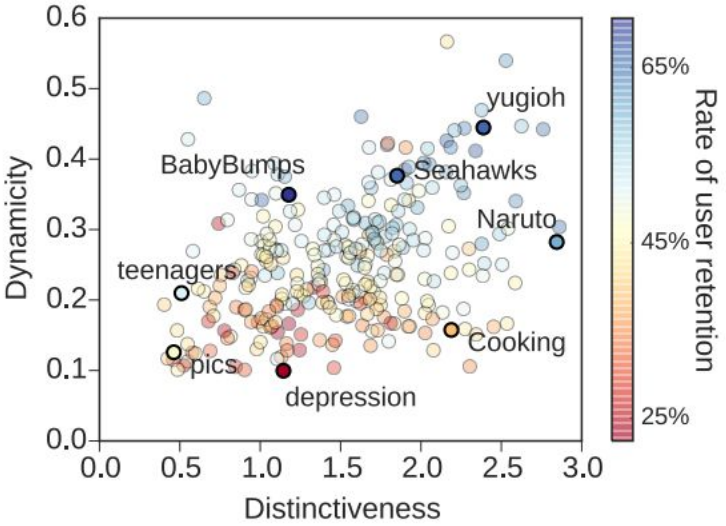
	generic	distinctive
dynamic	BabyBumps IAmA Libertarian australia	CollegeBasketball Seahawks formula1 yugioh
consistent	AdviceAnimals funny news pics	Cooking Guitar MakeupAddiction harrypotter

Other example:
Startrek vs. TheWalkingDead

Relating type to user engagement patterns

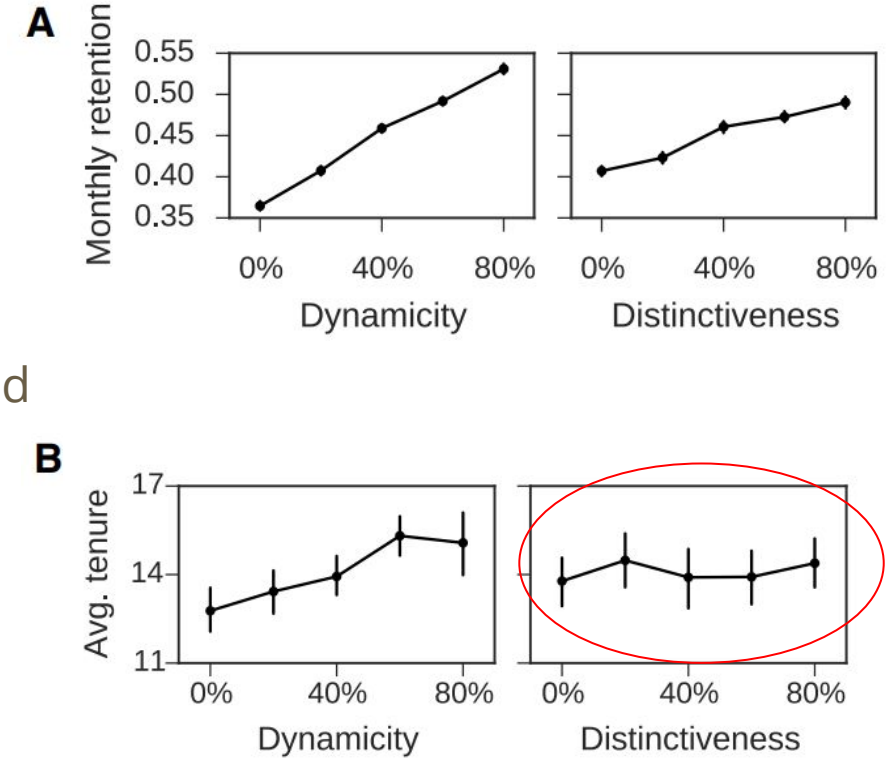
- Qualitatively satisfying groupings of communities
  - according to the nature of their collective identity

- **Is there an informative and highly predictive relationship between a community's position in this typology and its user engagement patterns?**
  - More Dynamic => higher rates of engagement
  - Distinctive ones => les universal, good in short-term not long-term



Community type and Monthly retention

- More Dynamic :
  - Higher rates of user retention
- More Distinctive:
  - ~ Higher rates of user retention
- Avg #months that user contributed to a community (last-first appearance)
  - distinctiveness does not correlate
  - with this longer-term variant of
  - user retention



Using Typology for prediction

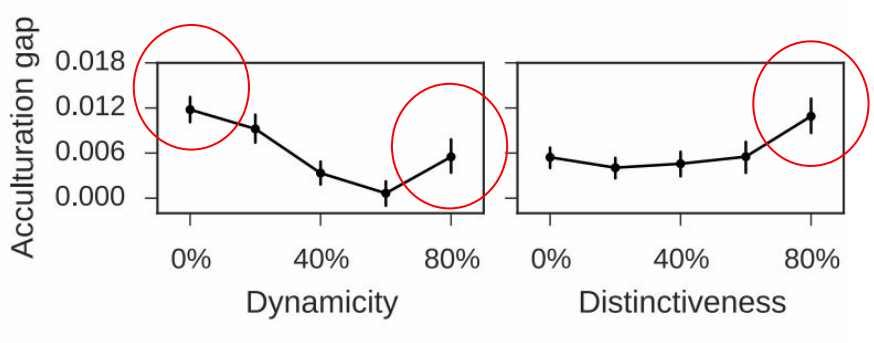
- This identity-based typology holds additional predictive value on top of strong baseline features
  - community-size (# contributing users)
  - activity levels (mean # contributions per user)

- Combining these features results in higher accuracy and lower errors.

Community identity and acculturation

- **Acculturation gap**:
  - compares the extent to which engaged vs. non-engaged users employ community-specific language
- Gap is most pronounced in
  - stable,
  - highly distinctive

  communities



Measuring the cross-entropy of bigrams for (users with at least 5 comments in respective community and month) vs. (those with only one)

Community identity and content affinity

- Strong correlation between **distinctiveness** and community **volatility gaps**
- In most distinctive communities
  - active users tend to write more volatile comments than outsiders
  - while across the most generic communities active users tend to write more stable comments

- They also find that in almost all communities,
  - **active users** tend to engage with more **community-specific content** than outsiders

The End