# \section{Introduction}

The past two decades have seen an explosion of large social computing projects, which encompass diverse communities of contributors.  Computer Mediated Communication (CMC) allows organizations to develop non-traditional management structures and for communities to collaborate both synchronously and asynchronously from geographically distant locations in order to create products and platforms of cultural, scientific, and technological significance, such as Wikipedia, the Zooniverse, and GNU/Linux.

One important research area examines the dynamics that govern relationships and social roles within these platforms and communities \cite{benkler_peer_2015}. Previous work confirms that structural difference exist within communities (e.g., communities show varying degrees of hierarchy and develop different processes for consensus building **{cite}**) and furthermore that community composition (e.g., editor experience and gender bias) impacts content \cite{keegan_editors_2012,wagner_its_2015}.  Yet the vast majority of these projects focus exclusively on culturally homogenous, English-based platforms.

## need a brief statement here on ML WP work to setup 2nd half of the arg

Our research fills this gap by testing culture's capacity to shape and structure sub-communities within a single platform.  We investigate differences and similarities surrounding collaboration dynamics within social computing systems that span linguistic and cultural boundaries. Ultimately, we hope to expand the broader understanding of peer-production and collective intelligence work beyond the bounds of purely English speaking groups, surfacing mechanisms that result in content variation and informing design solutions to better support global collaboration.

# \section{Background and Related Work}

## \subsection{Culture and Agency: Culture as a Structuring Force}

Culture and language shape societies in persistent, durable forms \cite{durkheim_elementary_2008}. The cultural contexts in which communities exist provide norms and expectations that structure the ways in which individuals interact, producing consistent patterns that can distinguish one group from another.  In other words, communities are culturally embedded, which creates similar relationships within groups and organizations, but structural and hierarchical differences between them.

Cultural structure can hypothetically influence diverse aspects of communities. A small body of qualitative and mixed methods work provides evidence that online spaces are culturally embedded across linguistic, temporal/spacial, religious, ideological dimensions \cite{orgad_cultural_2006}, and furthermore that language plays a particularly fundamental role in the construction and shaping of online communities \cite{danet_multilingual_2007}. However, while these studies further indicate that culture may structure collaboration dynamics, they tend to be highly specific to individual communities and therefore lack generalizability. One mixed methods study found significant differences in the ways that small/large, and east/west editor communities use talk pages, yet researchers sampled only thirty pages per language edition which could bias findings \cite{hara_cross-cultural_2010}.

% linguistic neighborhoods

Perhaps more convincingly, though little research explores culture's structuring role with respect to online collaboration dynamics, previous work indicates that culture indeed influences artifact production. Contrary to the global consensus of world knowledge hypothesis **{cite}**, encyclopedic content produced by Wikipedia editors does not cover topics universally due to differing degrees of cultural relevance among specific subjects \cite{hecht_tower_2010}. Services such as OMNIPEDIA and WikiBrain, which visualize these content asymmetries and biases across Wikipedia, reveal that a majority of subject matter is only available in specific language editions \cite{bao_omnipedia:_2012, sen_wikibrain:_2014}. Furthermore, varying gender biases differentiate Wikipedia language communities. Modeling approaches that analyze coverage, lexical, structural, visibility biases find that women and men portrayed differently across languages \cite{wagner_its_2015}. Yet while these studies indicate that substantial difference exist between encyclopedic *content* produced by different communities on a single platform, they focus on outputs instead of collaboration and production processes.

## \subsection{Collaboration and Coordination in English Speaking Communities}

Despite the relative lack of studies that explore collaboration dynamics at a multicultural level, these practices are well documented across a range of English speaking communities and social computing systems. Previous work reveals that editor coordination and conflict management shape English speaking Wikipedia article production, finding that levels implicit and explicit coordination are strongly associated with measures of article quality \cite{kittur_harnessing_2008}, but that coordination positively correlates with conflict \cite{kittur_beyond_2010}.

Both editor and group level characteristics also affect content outputs. English Wikipedia editors tend to focus their efforts on articles that cover similar topics rather than contributing to a number of different subject areas \cite{keegan_staying_2012}, and they return more frequently to edit newer articles. Furthermore, the success of interdependent tasks (such as improving

readability) requires small, concentrated groups of dedicated editors, while improving article coverage necessitates larger overall quantities of editors \cite{kittur_coordination_2009}. Yet, while more experienced editors and larger groups of editors tend to be more productive, the compatibility of editors within a group also affects productivity \cite{keegan_editors_2012}.

While these studies explore dynamics in English speaking projects and establish effective methods for exploring collaboration in large online communities, previous cross-cultural work indicates that findings may not generalize beyond English speaking communities. We aim to fill this gap in existing research with the following study by exploring collaboration practices within twenty five different language speaking communities. Specifically, our study combines and extends the two research areas described previously in order to extend our understanding of online collaboration beyond a small subset of cultures.

# \section{Data}

Current social computing literature addresses both collaboration within editor communities and content asymmetries across linguistically distinct groups, yet researchers have yet to conduct a broad comparative study that investigates the overlap between these two related areas. Wikipedia makes a compelling case study for hyperlingual peer production research due to its size, longevity, data access policies, and distribution across multiple language editions. Though researchers have raised valid concerns about the uniqueness of the Wikipedia project and the generalizability of Wikipedia based results to other peer production and collective action projects **{cite}**, we nevertheless believe Wikipedia provides the best dataset for our work due to its parallel yet distinct language communities.

Our research relies on Wikipedia revision histories from article and talk pages. These categories (article and talk) are known as *namespaces* within the wiki community. Analysis will leverage each namespace as follows:

## \subsection{Article Pages}

Article Pages are the default view of Wikipedia and contain encyclopedic content. A single article corresponds to a unique concept---for example ``language'' or ``book''. Though articles make up the bulk of Wikipedia's forward-facing pages and article histories provide a time-stamped change-log of content additions and edits, they only represent the single, end product viewpoint of editor collaboration. Though we primarily analyze talk pages, article pages provide control variables in our models.

## \subsection{Talk Pages}

Talk Pages contain backchannel conversations that potentially surface a more nuanced view of editor collaboration. Each article links to an associated Talk Page in which editors discuss necessary additions or edits to the main article.  These conversations follow the same structural conventions as articles---editors add or change content under sections which link to a specific, unique topic---and can therefore be analyzed using similar techniques (discussed in the following section).  The advantage of talk pages is that they capture back-channel conversation, which theoretically shape content presented in associated article pages.  Like article pages, WikiMedia provides a full talk page edit history that contains timestamps and editor usernames or IPs.

## \subsection{processing}

Our collection of article and talk pages currently covers the 25 distinct language editions used by Hecht \& Gergle \cite{hecht_tower_2010}, and includes the complete revision history as well as meta-data for each revision.  Between November 2015 and May 2016 we downloaded the entirety of this Wikipedia dataset as xml data-dumps, which are supplied by the WikiMedia Foundation for archival and research purposes.

Uncompressed, this dataset would require hundreds of terabytes **(check this)** of storage (the complete revision history of English Wikipedia alone is **(num)**).  We therefore stored the xml-dumps as compressed archives, and dynamically uncompressed and re-archived each as necessary while producing qualitative measures for our models.  All data collection, processing, and analysis—with the exception of qualitative coding—was tested on a local dedicated research server and executed on a large computing cluster.  We used a series of custom python scripts and the Mediawiki-Utilities **{cite}** python library to parse, clean, and processes each xml dump.

% we removed bot edits

# \section{Analytical Approach}

## \subsection{Quantitative Measures}

Our analytical approach leverages mixed qualitative and quantitative research methodologies. Following approaches common to online collaboration research, we first reduced each article and talk page to edit counts, which captured the following page level measures:

*Num_talk_edits:* The number of edits to a particular talk page
*Num_article_edits:* The number of edits to a particular talk page
*Unique_talk_authors:* The number of unique contributors to a specific talk page

*Unique_article_authors:* The number of unique contributors to a specific article page
*Talk_age:* The elapsed time between the first and final edit to a specific talk page
*Article_age:* The elapsed time between the first and final edit to a specific article page
*Namespace:* Indicates whether the page is an article or a talk page
*Language:* The language edition to which the page belongs.

Each article/talk pair has a unique title, which we used to match talk pages with corresponding articles. Not all article pages possess a talk page (talk pages are not automatically created for each article and frequently small articles do not require talk), so after matching we discarded all unpaired articles. Since this study examines coordination and collaboration, we also dropped all talk/article pairs which contained a page with only a single edit. Overall this accounted for **(num)** articles and **(num)** talk pages within our dataset, broken down by language in **(figure x).**

## \subsection{Descriptive Statistics and Models}

We then calculated basic page level descriptive statistics for each language edition using python's pandas **{cite}** library. Additionally, we plotted the distributions of page level edit counts for both article and talk pages and calculated the skew of each.

Creating interpretable statistical models of Wikipedia edit counts poses a number of methodological problems due to both size and distribution. Following **{cite}**, we used a negative binomial distribution, which adjusts estimates for long tailed and over dispersed data **{cite}**. The negative binomial distribution extends the relatively common Poisson distribution, but adds an additional parameter in order to model overdispersion. Due to convergence issues with several R packages, we ultimately used the STATA implementation to generate parameter estimates. Additionally, we dropped a single talk/article pair of over 50,000 edits in order to make our models converge.

## \subsection{Qualitative Coding}

% something about Gieger + trace ethnography, spot checking, etc.

# \section{findings}

Preliminary analysis indicates that collaboration practices---specifically talk page edits---do indeed vary by language community. Specifically, though the quantity of edits to a specific talk page is somewhat correlated with other variables including age and number of editors, the language edition to which the page also appears to play an important role.

In order to test the role of language, we built three separate models. The first model predicts talk page edits based on the number of editors who have contributed to that talk page, and the page's age in years (*num_talk_edits ~ unique_talk_authors + talk_age*). For an average talk

page, a one editor increase results in an 8.3 percent increase in edits, and each year of age results in a 7.9 percent edit increase.

Our second model controls for variables associated with a talk page's corresponding article pag, such as article edits and article age (*num_talk_edits ~ unique_talk_authors + talk_age + num_article_edits + article_age*). We omitted the *num_article_editors* due to correlation of .95 with *num_article_edits*. For an average talk page, a one year increase in the corresponding article's age corresponds to a 2.4 percent increase in talk edits, though a single article edit increase corresponds to only a .02 percent increase. Additionally, adding article level controls decreased the influence of *unique_talk_authors* and *talk_age* to 7.7 and 6.2 percent respectively.

The final model demonstrates the association between language and talk page edits. In addition to the controls listed above, we added 25 language dummy variables, one for each language (*num_talk_edits ~ unique_talk_authors + talk_age + num_article_edits + article_age + lang*). The coefficients on each of these variables represent the expected change in number of talk edits due only to the language edition. We calculated all coefficient estimates with respect to English Wikipedia.

%<Insert descriptive stats table comparing wikis>
%<insert regression table>

% Insert paragraph that just summarizes the models

Hebrew and Vietnamese represent the two extreme cases, where an average talk page in Hebrew Wikipedia will contain 62.6 percent more edits than an English talk page, while a talk page in Vietnamese Wikipedia will contain 31.2 percent fewer edits. Somewhat surprisingly, though English Wikipedia supports the largest, oldest, and most established editor community, it falls near the middle of the distribution; 10 language editions on average contain less talk page activity, while 15 contain more. Of those editions that contain more talk page activity, 6 contain over 25 percent more (Hebrew, Japanese, Finnish, Czech, Swedish, and German), while only 2 contain 25 percent less (Vietnamese and Portuguese).