University of Pennsylvania
Dept. of Computer and Information Science
Philadelphia, Pennsylvania, USA

In Partial Fulfillment of the WPEII Requirement

Calculating and Presenting Trust in Collaborative Content

Author: Andrew G. West

Committee Chair: Sampath Kannan
Committee Member: Andreas Haeberlen
Committee Member: Boon Thau Loo

October 2010

Contents

Abstract

Collaborative functionality is increasingly prevalent in Internet applications. Such functionality permits individuals to add – and sometimes modify – web content, often with minimal barriers to entry. Ideally, large bodies of knowledge can be amassed and shared in this manner. However, such software also provides a medium for biased individuals, spammers, and nefarious persons to operate. By computing trust/reputation for participating agents and/or the content they generate, one can identify quality contributions.

In this work, we survey the state-of-the-art for calculating trust in collaborative content. In particular, we examine four proposals from literature based on: (1) content persistence, (2) natural-language processing, (3) metadata properties, and (4) incoming link quantity. Though each technique can be applied broadly, Wikipedia provides a focal point for discussion. Finally, having critiqued how trust values are calculated, we analyze how the presentation of these values can benefit end-users and application security.

1 Introduction

Collaborative functionality has become a pervasive part of the Web browsing experience. Topical forums, blog/article comments, and open-source software development are all examples of *collaborative applications* – those that enable a community of end-users to interact *or* cooperate towards a common goal. The most fully-featured of such models is the *wiki* [43], a web application that enables users to create, add, and delete from an inter-linked content network. On the assumption that all collaborative systems are a reduction from the *wiki* model (see Sec. 2.1), we use it as the basis for discussion.

No doubt, the collaborative encyclopedia Wikipedia [10] is the canonical example of a *wiki* environment. Wikipedia also provides ample evidence of the abuses possible in such settings. Nearly 9% of all edits to Wikipedia are *reverts* (undos), which are often used to repair gross incompetence [17]. Although there is oft-cited evidence defending the accuracy of Wikipedia articles [31], it is negative incidents that tend to dominate external perception. For example, Wikipedia has contained death hoaxes [49], reported potentially libelous content [50], been a distribution point for malware [39], and Wikipedia-sourced information has led to errors in traditional media [56].

While these examples erode Wikipedia's reputation (and perhaps pose a legal threat), far more damaging scenarios can be imagined. For example,

imagine an article (on Wikipedia, or otherwise) which has been vandalized to distribute false medical advice. Alternatively, consider a rogue employee altering information on Intellipedia (a *wiki* for U.S. intelligence agencies), on which military decisions may be based.

It is not just *possible* to attack collaborative applications, but certain characteristics of the model make it *advantageous* to attackers: (1) Content authors have access to a large readership that they did not have to accrue[1]. (2) A single edit can be viewed a limitless number of times (contrast this to the one-to-one model of sending email spam). (3) Anonymous editing means 'real-world' reputations are not at stake. (4) The open-source nature of much *wiki* software makes security functionality transparent.

Combining these vulnerabilities and the discussed prior incidents, the need trust metrics in the collaborative domain should be apparent. Indeed, the need to identify trustworthy agents/content has been the subject of many academic writings and on-wiki applications. In the remainder of this paper we survey those approaches, which we broadly classify into four distinct approaches. Where appropriate, we highlight both the paper that introduced the technique (often not in the collaborative domain), and the paper that has best applied it in a collaborative setting (*i.e.,* Wikipedia):

1. Content Persistence: Building on the work of [63], Adler *et al.* [19, 20] propose a system whereby the persistence of an author's content determines his/her reputation. In turn, author reputation can speak to the quality/trust of new content authored by that contributor.

2. Natural-Language Processing (NLP): Akin to the use of NLP in email spam detection [55], the proposal of Wang *et al.* [60] uses language features to distinguish damaging edits from quality ones.

3. Metadata Properties: Just as the SNARE system [33] did for email spam, Stvilia *et al.* [58] identify poor contributions by looking at the *metadata* for an edit – properties unrelated to the linguistics of the content (*e.g.,* article size, time-stamp, account age, *etc.*).

4. Incoming Link Quantity: Based on well-known algorithms for search-engine ranking [40, 48], the work of McGuinness *et al.* [45] proposes that pages with a large number of incoming links (internal or external of the *wiki*) are likely to be reliable resources.

[1] As of this writing, the English Wikipedia averages 7 billion views/month [17].

After describing each of these approaches in greater depth, discussion will shift to their relative merits. That is, how do the systems perform? How robust is each to evasion? How do they compare in terms of computational speed? How are new users initialized?

Calculating predictive and/or representative trust values is meaningless unless they are effectively conveyed to the end-user. Thus, we review proposals on how trust should be presented in collaborative applications. The combination of effective trust calculation and presentation holds enormous potential for collaborative applications of the future.

The remainder of this work is structured as follows. In Sec. 2 we establish the terminology of collaborative systems, discuss the formalization of trust, and discuss the granularity of trust computation. Then, in Sec. 3 we describe the varied approaches to trust computation and their relative strengths and weaknesses. Sec. 4 focuses on how these computed values can be presented to benefit both end users and system security. Finally, concluding remarks are made in Sec. 5.

2 Background & Terminology

In this section, we standardize terminology for the remainder of this work. Further, we attempt to define the meaning of trust in a collaborative environment and claim that no matter for what entity trust is calculated (article, user, edit, *etc.*), the notion is transferrable to other participating entities.

2.1 Defining a Collaborative System

Put simply, a *collaborative system* is one in which two or more *users* or *contributors* operate in a centralized shared space to achieve a common goal. Taken as a whole, the user-base is often referred to as a *community*. Among the most distinguishing factors between collaborative systems is (1) the *accessibility* of the collaborative tool to casual users, and (2) the extent to to which read/write/create/delete *permissions* are extended to the community.

We are primarily concerned with systems which are freely accessible and encouraging of widespread participation. For example, many *wikis* (Wikipedia included) have no barrier to entry and allow anonymous users to contribute freely. Similarly, applications that require a CAPTCHA solve or free registration are not imposing a significant burden. On the other hand, corporate

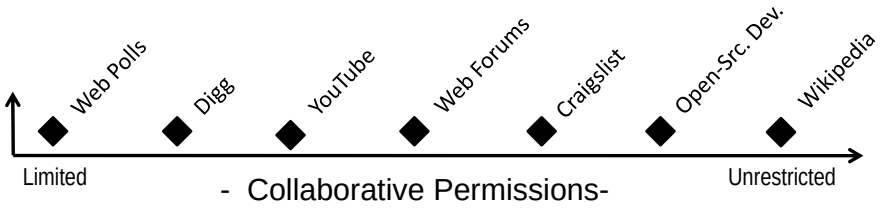

Figure 1: Various Accessible Collaborative Applications

Wikis and repositories have high barriers to entry (you must be an employee) and the lesser degree of anonymity likely mitigates most malicious behavior.

Even among accessible systems, the user-facing permissions vary dramatically. One of the most constrained examples of a collaborative system is a "web poll", where users can select among pre-defined options and submit a response which is stored and displayed in aggregate (graph) fashion. A more permissive example is "blog comments", where readers can append content of their choosing on existing posts. At the extreme of this continuum lies the *wiki* [43] philosophy, which in its purest[2] form gives its users unrestricted read/write/create/delete permissions over all content. Figure 1 visualizes this continuum, orienting some well known collaborative use-cases.

Given that the *wiki* model extends *all* file permissions to *all* participants, it is reasonable to assume that all other collaborative systems must operate using a sub-set of these permissions and accessibility. We believe this justifies our decision to concentrate discussion at the *wiki*-level, as it is the most generalized of such models. Indeed, many of the techniques herein can be, and have been, applied in more restrictive collaborative settings.

In particular, Wikipedia [10] is the most well-known use-case of the *wiki* philosophy. Further, it has become a *de facto* standard in the evaluation of collaborative trust systems (and collaborative work in general). For these reasons, our discussion moving forward will focus heavily on Wikipedia.

2.2 Wiki(pedia) Terminology

Given our focus on *wiki* environments, it is helpful to standardize terminology. A *wiki* consists a set of content *pages*, *articles*, or *documents*. Content

[2]Wikipedia is not a *wiki* in the purest sense. The realities of operating a web presence of that magnitude have led to the installation of minimal protections.

evolves through a series of *revisions* or *edits*, which taken in series form a *version history*, $R = \{r_0, r_1, r_2 \dots r_n\}$. Though it is possible to browse a page's version history, it is the most recent edit, r_n, which is displayed by default. A special form of edit called a *revert* or *undo* creates a new version, but simply duplicates the content of a previous one. Reverts are interesting because they are often used to remove content which is deemed destructive.

An edit is made by exactly one *editor*, *contributor*, or *author*. Authors may be assigned persistent identifiers that allow their contributions to be tracked though time. Alternatively, some systems permit authors to edit in a more anonymous and transient fashion (see Sec. 3.2.1).

Individual pages within a *wiki* can be interconnected via hyperlinks, and such links are termed *internal links* or *wikilinks*. These should be distinguished from hyperlinks which lead users to destinations outside the *wiki* environment, known as *external links*.

Often, as *wikis* and their user-bases evolve, so does a specialized terminology. For Wikipedia, readers may find the following glossary [13] helpful.

2.3 Defining Collaborative Trust

2.3.1 Absence of Formal Definition

As Jøsang [37] notes, the meaning of *trust* varies dramatically throughout research literature. The collaborative domain is no exception. Such writings often handle trust generically; giving readers little insight into precisely what properties the calculated values are meant to quantify.

From our comparative standpoint this is unsatisfactory, but a rigorous and objective definition for *trust* is seemingly hard to construct (as we discuss in Sec. 2.3.3). Even if such a definition existed it would do not aide our analysis of existing systems (which would have been created without respect for that definition). Moving forward, we emphasize *how* the systems operate and consider the subtleties of *what* the systems are calculating to be beyond the scope of this work. Most importantly, comparison between these systems remains possible because of their generic evaluation criteria.

2.3.2 Basis for Evaluation

Because definitions of *trust* are so varied and often include some degree of subjectivity, it becomes difficult to establish ground-truth for the evaluation

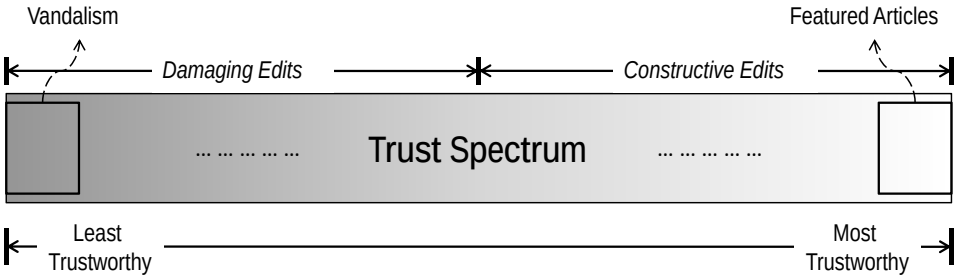

Figure 2: Collaborative Trust Spectrum

of trust management systems. As a result, performance measurements either, (1) concentrate on the most objective subset of the trust spectrum (see Fig. 2), or (2) divide the trust spectrum at coarse granularity.

For Wikipedia-based analysis, vandalism detection is the most prominent example of the first technique. *Vandalism* is defined to be any edit which exhibits ill-intentions or gross negligence on the part of an editor. Vandalism often manifests itself as non-sense, obscenity, or spam [53] and occupies the extreme-left of the trust spectrum. Thus, vandalism is the least trustworthy of all content, and perhaps the easiest to label as such[3]. As a result, it is easy to amass large labeled corpora for evaluation purposes [52].

The second evaluation strategy divides the trust spectrum at coarse granularity. On Wikipedia, a set of community-labeled pages known as *featured articles* are generally used as examples of trustworthiness. The subjective "featured-article criteria" [12] are used as a guideline for labeling. For evaluative purposes, featured articles are generally contrasted against the remainder of articles, or a subset of articles known to be of poor quality.

While these two evaluation techniques represent the current state-of-the-art, they are less than ideal. First, vandalism detectors operate on a subset of the trust problem, so it remains to be seen if the same metrics are meaningful at the far-right of the trust spectrum. That is, can the same criteria be applied to distinguish mediocre edits from quality ones? Indeed, it would seem a holistic measurement of trust might be more complex.

Second, treating trust as a two-class problem seems inappropriate (as

[3]Identifying the *most* trusted content would be considerably more difficult. Not only is there ambiguity and subjectivity as to what constitutes *trust* – but merely gauging an edit's factuality could require subject experts.

coarse-granularity approaches tend to employ) as it captures no subtleties. For example, it is unsurprising that good articles are usually longer than poor ones. However, article length may be a poor comparator among a set of reasonable articles. Lastly, both approaches are able to rely on community-based labeling, allowing author's to side-step the need for precise definitions regarding how content should be tagged (and thus, what constitutes *trust*).

2.3.3 Discussion on Collaborative Trust

Despite the fact a precise definition for collaborative trust is unnecessary for purposes of evaluation or our comparative discussion – there is still merit in examining how such a definition might proceed.

One may be tempted to think that trust may be equivalent to a document's "information quality" (IQ) [59]. Information quality metrics such as validity, currency, objectivity, comprehensiveness, *etc.* are a common method of evaluating information systems (including Wikipedia [58], see Sec. 3.1.3).

However, we believe such analysis fails to take the subjectivity of online environments into account. As a counter-example, consider Encyclopædia Dramatica [3], a *wiki* which parodies Wikipedia by encouraging the contribution of biased and offensive content[4]. The most 'celebrated' of such content would surely have poor IQ metrics. As a less extreme example, consider that Wikipedia's guidelines set forth the objective to create an online encyclopedia. Thus, even an article which is completely, factual, and objective (*i.e.,* strong IQ) – may be deemed untrustworthy (and deleted) if the subject matter fails to meet notoriety requirements. Thus, we propose ...

> *Content trust must be measured through the subjective lens of the community consensus on which it resides. It will be the case that trustworthy content exists precisely where that content meets the expectations set forth by the project community.*

Of course, this depends on how the community expectations are defined. Wikipedia does this quite explicitly with policy pages and the featured article criteria [12]. For other communities, it may be possible to learn of expectations implicitly, by observing common editing practices. Ideally, communities should (1) make such expectations explicit, and (2) define them in the most

[4]This site may be inappropriate to view in workplaces or educational settings.

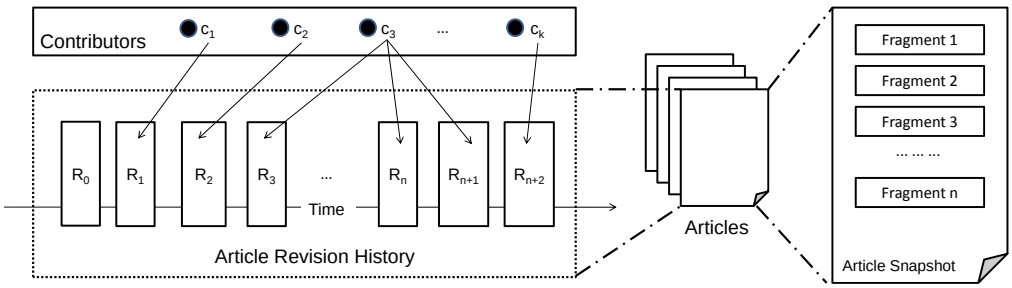

Figure 3: Relationship Between Wiki Entities

objective terms possible. Adherence to these suggestions would simplify both the task of editors, as well as that of the researchers who study them.

In many reasonable *wiki* environments, community expectations may define 'trustworthy' synonymously with 'quality.' However, respecting our subjective definition, we will call edits that are viewed favorably as *constructive* or *value-adding*, and poorly viewed edits will be termed *damaging*.

Many of the systems discussed in the coming pages were designed specifically for use on Wikipedia. Therefore they implicitly incorporate the expectations of that project. It remains to be seen if such approaches can be generalized for use in alternative *wiki* settings (see Sec. 3.2.5).

2.4 On the Associativity of Trust

The methodologies we will examine in the coming section calculate trust values for either (1) articles, (2) article fragments, (3) revisions, or (4) contributors. We assume that these entities have an associative trust relationship. That is, if one has trust values for any one of these sets, than this is sufficient to calculate trust values for the other three types[5]. For example, the trust values of all edits made by a particular author should speak to the trust of that author. Similarly, the trust of all text fragments of an article should speak to the reputation of that article. Thus, all collaborative trust systems are calculating at the same granularity and can be treated as comparable. Figure 3 visualizes the relationship between these different entities.

[5]While these associative relationships can be defined, different trust systems generally excel when operating at a particular granularity (as detailed later in Table 5).

Approach		Strength	Weakness	Paper
Content-persist		Implicit feedback mechanism holds f-back providers accountable	Difficulty with Sybil and new users. Reliant on hindsight	Adler [19, 20]
NLP	Lexical	Regexps easy to implement, modify, and understand	Evadable by obfuscating or avoiding poor language	Wang [60]
	n-gram	Find unusual or bad text w/o manual rules	Processing topic-specific corpora is CPU expensive	
Metadata-based		Size/diversity of available feature space	Properties are "a level removed" from content	Stvilia [58]
Citation-based		Calculation breadth makes evasion difficult	Unclear if citation action actually speaks to article trust	McGuinness [45]

Table 1: Signature strengths and weaknesses of approaches

What is not precisely defined are the mathematical functions that define these associative relationships. Occasionally, systems define these in an application-specific manner. On the whole, we consider this to be outside the scope of this work and an open research question.

3 Collaborative Trust Algorithms

3.1 Introduction to Algorithm Function

In this section, we feature a characteristic paper for each trust calculation technique. For each approach, we intuitively describe the algorithm's operation and describe related works. Table 1 summarizes the characteristic strengths and weaknesses of each approach, independent of the other techniques (and later, Figure 9 summarizes the related works and research timeline for each approach).

3.1.1 Content-driven Trust

Approach: As detailed by Adler *et al.* [19, 20], content-persistent trust is built on the intuition that the survival/removal/restoration of text fragments in subsequent revisions speaks to the trust of that fragment and to the reputation of its author. Content which survives future revisions, especially those of reputable authors, is likely to be trustworthy. Content which is removed but eventually restored is also trustworthy, but content which remains deleted speaks poorly of that content and its contributor.

Two quantities are used to define the notion of persistence. First, *text-life* is the percentage of space-delimited words added in some revision, r_i, which persist after a subsequent edit, r_j. The second is *edit-distance*, which measures the extent to which reorganization and deletions are preserved. The authors' develop a specialized `diff` algorithm to quantify the latter quantity.

Assume author A has made edit r_n on some article, and some time later, author B edits the same article, committing version r_{n+1}. At this point, the reputation of author A can be updated proportional to four factors: (1) the size of A's contribution, (2) the text-life of r_n relative to r_{n+1}, (3) the edit-distance of r_n relative to r_{n+1}, and (4) the reputation of B. The reputation of A will be further updated at each subsequent edit until r_{n+10} is reached. The reputation of A speaks directly to the trustworthiness of A's content, which is especially useful in judging new contributions of A which are yet to be vetted by subsequent editors.

Figure 4 helps exemplify the content-persistence algorithm. Assume authors A_1, A_2, and A_3 are equally trusted, and author A_1 initializes the "Benjamin Franklin" article with content to form version V_1. The actions of editor A_2 in version V_2 challenge the veracity of A_1, since he modifies content from V_1. However, when A_3 restores the content of A_1/V_1, it is A_2's reputation which is punished. When V_4 is committed, A_2's reputation is further reduced, and the statement "Mr. Franklin flew a kite" gains reputation, as well as the authors who endorsed this view (A_1 and A_3) – and this process would continue to some specified depth (Adler uses $depth = 10$).

Related Works: Adler's system is both a formalization and refinement upon the informal proposal made in [26] by Cross, which suggests that text-age may be indicative of fragment trust. Whereas Cross would treat restored text as new and potentially untrustworthy, Adler investigates the transience of content through greater revision depth.

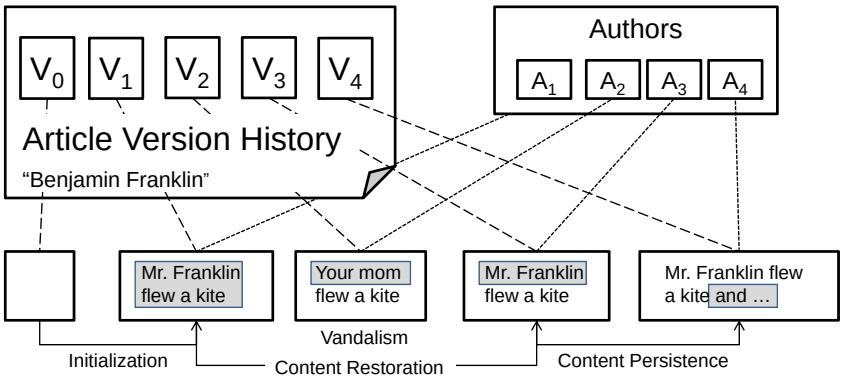

Figure 4: Example Content-Persistence Calculation

The system most related to Adler's is that of Zeng *et al.* [63] who used Dynamic Bayesian networks and the Beta probability distribution to model article quality. Zeng's system takes both author reputation and `diff` magnitude as inputs when calculating article trust. Whereas Adler computes predictive author reputation, Zeng uses pre-defined *roles* (*e.g.,* administrator, registered, anonymous, *etc.*) as predictors of author behavior.

Also similar is Wöhner *et al.* [62], which measure content persistence and transience rates throughout an article's lifespan. They find that quality articles are defined by a stage of high editing 'intensity', whereas low quality articles tend to have little of their initial content modified as they mature.

The notion of author reputation was also investigated by West *et al.* [61]. Rather than doing fine-grained content analysis of Adler, West detects an administrative form of revert called *rollback* to negatively impact the reputations of offending editors. Reputations improve only via the passage of time and this lack of normalization is problematic because rarely-erroneous prolific editors may appear identical to dedicated but small-scale vandals.

Two other systems based on content-persistence include [35, 44]. However, these attempts are under-developed or under-evaluated in comparison to the described efforts.

Live Implementation: The proposal of Adler has been implemented as a live Wikipedia tool, WikiTrust [18]. WikiTrust colors text fragments to display the associated trust values (see Sec. 4.2.1).

3.1.2 NLP-based Trust

Approach: Distinct from content-persistence (Sec. 3.1.1) which treats words as meaningless units of content, natural-language processing (NLP) techniques analyze the language properties of tokens. The techniques are varied; from simple text properties (*e.g.,* the prevalence of capital letters), obscenity detection (via regular expressions), to text similarity and predictability (n-gram analysis). We choose the recent work of Wang *et al.* [60] to be representative of this domain due its breadth of techniques.

Wang (and practically all NLP-based works) produce a feature-vector over which traditional machine-learning techniques are applied. In particular, Wang *et al.* divide their feature-set into three different NLP-driven categories: (1) lexical, (2) semantic, and (3) syntactic.

Lexical features are straightforward and are generally implemented via regular expressions. For all content added in a revision Wang implements a measure of, (i) vulgarity, (ii) slang (*e.g.,* 'LOL' or 'p0wned' – phrases which are not obscene, but improper in formal English), and (iii) improper punctuation (*e.g.,* the repetitive usage of question or exclamation marks).

The *syntactic* and *semantic* categories are more complex. For syntactic analysis, Wang performs n-gram analysis using only part-of-speech (POS) tags. That is, using some corpus (general or topic-specific) one computes the probability of all POS sequences of length n. Then, when an edit is made, the probabilities of new POS sequences are calculated. Improbable POS sequences are likely indicative of a damaging edit. Wang's semantic analysis also uses n-gram analysis but uses unique words instead of POS tags.

Figure 5 shows an example analysis using semantic unigrams (*i.e.,* $n = 1$). Related sources are amassed to build a dictionary of words common in discussion of the article under investigation, "Benjamin Franklin." When words added to the article elicit a high "surprise factor" (*i.e.,* have not been seen in the corpus), there is good probability of suspicious activity. Indeed, Ben Franklin never flew a jet, and the revision is vandalism.

Related Works: The work of Wang is recent to this writing and incorporates many ideas from earlier literature. Many such works investigated the predictive nature of n-gram analysis. One of the first was Smets *et al.* [57], utilizing Bayesian analysis (initially shown useful in email spam detection [55]) and Probabilistic Sequence Modeling. Similarly, [23] used a generic predictive analysis, while Itakure *et al.* [34] leveraged dynamic Markov com-

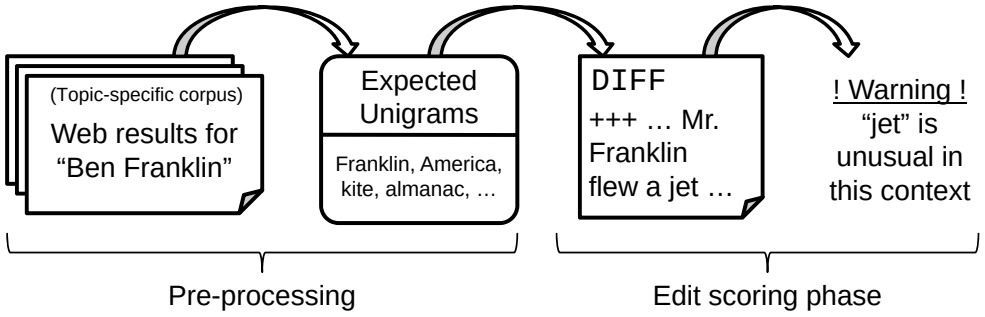

Figure 5: Example NLP Semantic n-gram Calculation

pression. While different in technique, these techniques are calculate roughly equivalent probabilities. However, the work of Wang is unique in that probabilities are generated from web-based corpora (*i.e.,* the top k search-engine results for a topic), whereas earlier literature used only the (more narrow) Wikipedia article itself or a generalized corpus.

Distinct from predictive techniques are those of Potthast *et al.* [51] which tend to focus on aggregate-count measures. For example, Potthast includes simplistic features such as (i) ratio of upper-case characters, (ii), longest word length, and (iii) pronoun frequency. Along the same lines, Rassbach *et al.* [54] use an un-described set of "about 50 features" from an NLP toolkit.

Also in the NLP realm would be the 'readability' measures (*e.g.,* Flesch-Kincaid, SMOG) incorporated into some trust systems [54, 58]. Though collaborative literature provides little insight regarding their function or usefulness, these systems produce a measure of text complexity by examining sentence lengths and syllable counts.

Live Implementation: NLP techniques are being applied in real-time on Wikipedia by an autonomous script called ClueBot [1], which automatically reverts trivially offensive edits. Due to a low tolerance for false-positives, ClueBot operates using a conservative and manually-authored set of regular expressions. ClueBot has been well studied [30, 57, 60] and exemplifies that lexical measures need not be strictly punitive. For example, regexps capturing advanced *wiki*-markup can increase edit trust.

3.1.3 Metadata-based Trust

Approach: If we consider article versions to be the *data* in a *wiki* system, *metadata* is then any property which describes that data[6]. We divide metadata into two sets: content-exclusive and content-inclusive.

Content-exclusive properties consider only descriptors external of article text. For example, each edit has a: (1) time-stamp, (2) editor, (3) article title, and (4) edit summary[7]. These can then be aggregated (for example, to compute the number of unique editors in an article's history), or combined with external information (on or off the *wiki*).

Meanwhile, *content-inclusive* measures permit summarization of the article or `diff` text. For example, this could be a measure of article length or the number of images in an article. Indeed, some degree of text-parsing would be required to extract these properties. Thus, we believe such properties may verge on being lexical NLP ones (like those of Potthast [51]). In general, we prefer language-driven features of this kind to be classified in the NLP domain and structurally-driven ones considered metadata.

Regardless, systems of this kind proceed by identifying multiple metadata-based indicators and producing predictive measures via machine-learning. Table 2 lists several example features of each type. Incorporating many of these features is the work of Stvilia *et al.* [58], which we choose to be representative of metadata-based approaches.

Rather than simply identifying metadata indicators, Stvilia takes an information quality (IQ) approach. IQ metrics [59] are properties like completeness, informativeness, consistency, and currency which generally define document quality (even outside of collaborative environments [64]). Stvilia's contribution is the quantification of these metrics for Wikipedia via the use of metadata features. For example, a measure of *completeness* considers the article length and the number of internal links. *Consistency* considers an article's age and the percentage of edits made by administrators. This IQ-based approach seems a more intuitive and elegant use of metadata than simply pushing raw-features to a machine-learning framework.

[6]By definition, properties we have separated out as entirely different techniques (*e.g.,* content persistence) could also be considered content-inclusive metadata. For consistency, reader's that believe these categories to be in conflict should consider only content-exclusive properties to be part of a metadata-based approach.

[7]An optional text field where an editor can briefly summarize changes.

Content-Exclusive Features	
Editor	**Time-stamp**
· Anonymous/registered	· Local time-of-day
· Time since first edit	· Local day-of-week
· User edit count	· Time since article edited
Article	**Revision Summary**
· Num. edits in history	· Comment length
· Article age	· If edit marked 'minor'
Content-Inclusive Features	
· Article length	· Revision `diff` size
· Num. external links	· Num. images

Table 2: Example Metadata Features [22, 51, 58, 61]

Related Works: The work most similar to Stvilia's is that of Dondio *et al.* [27]. Dondio begins by formally modeling the Wikipedia infrastructure and identifying ten "propositions about trustworthiness of articles" which are essentially IQ metrics. However, only two metrics are developed (fulfilling three of the propositions), *leadership* and *stability*. These "domain-specific expertise" metrics are shown to marginally improve on cluster analysis over 13 raw metadata features (*e.g.,* article length, number of images).

Meanwhile, inspired by the use of metadata to combat email spam [33], West *et al.* [61] concentrate on a narrow set of content-exclusive metadata features based on spatio-temporal properties. Simple properties include the time when an edit was made, the length of the revision comment, and how long the editor had been a community participant. More novel are reputations generated from metadata-driven detection of revert actions. Article and author reputations are straightforward, but *spatial reputations* for topical-categories and geographical regions are novel in their ability to have predictive measures available for new entities.

Almost comical compared to the complexity of these approaches, Blumenstock [22] claims that a single metric – word count – is the best indicator of article quality and significantly outperforms other discussed strategies.

Live Implementation: Metadata properties are being used to evaluate Wikipedia edits in a live fashion. The STiki anti-vandalism tool [8] is built on the logic of West's approach. It calculates trust scores which are used to prioritize human-search for damaging edits (see Sec. 4.2.2).

3.1.4 Citation-based Trust

Approach: Borrowing from citation-based algorithms commonly used in search-engine retrieval ranking such as HITS [40] and PageRank [48], McGuinness *et al.* [45] propose a *link-ratio* algorithm.

First, consider an article, a_n on Wikipedia (*e.g.,* "Benjamin Franklin"). The title of a_n can then be treated as an *index term* and full-text search can be conducted on all other *wiki* articles (*i.e.,* $\forall a_i, i \neq n$), counting the number of occurrences of that term (*e.g.,* articles like "Philadelphia" or "bifocals" are likely to have occurrences of "Benjamin Franklin").

Each of these occurrences are then labeled. Occurrences formatted to be internal wiki-links (*i.e.,* the index term is a hyperlink to the matching article) are termed *linked*, whereas occurrences where this is not the case (*i.e.,* the term appears as plain-text) are *non-linked*. The ratio of linked occurrences to all occurrences is the *link-ratio*, the metric of interest. McGuinness argues that high link-ratios are indicative of trusted articles, as the decision to cite another article is an implicit recommendation of that article's content.

An example of McGuinness' algorithm is visualized in Figure 6 (using our "Benjamin Franklin" example) – note that the `[[...]]` syntax is common *wiki* markup for internal links. To give some idea of the scale at which such algorithms operate, the actual "Ben Franklin" article has over 4000 incoming citations as of this writing.

Related Works: In the course of their evaluation, McGuinness *et al.* compared their link-ratio algorithm to results using the PageRank algorithm [48]. Earlier, Bellomi *et al.* [21] performed internal network analysis using both the PageRank and HITS [40] algorithms. The major difference between the link-ratio and search-inspired strategies is the extent of normalization.

For example, if an index term appears just once in the full-text of the *wiki*, and that once instance is linked, than the term will have a perfect link-ratio. Thus, to increase an article's trust value, one need only convert existing plain text references to linked ones. In contrast, PageRank and HITS perform more complex (non-normalized) graph analysis.

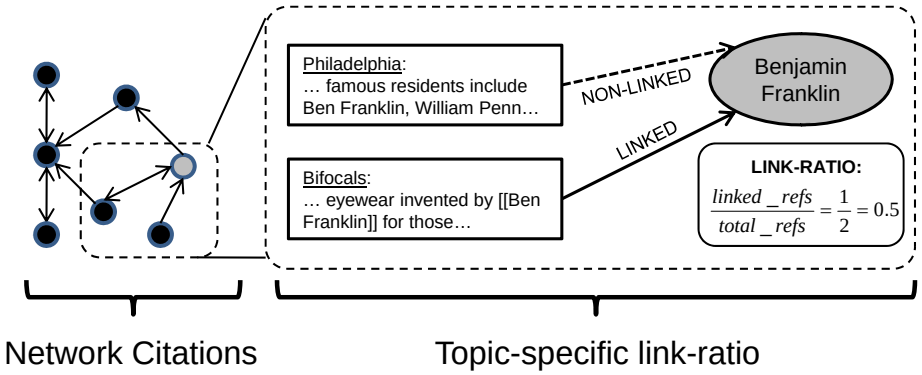

Figure 6: Example Link-Ratio Calculation

Live Implementation: To the best of our knowledge, there is no live implementation calculating citation-based trust for Wikipedia. However, Google's browser toolbar [5] exposes the PageRank values calculated by the search-engine provider, which could be interpreted comparatively.

Wikipedia does identify *orphaned* articles – those with few or no incoming links. While Wikipedia provides such lists [7] to encourage the strengthening of network connectivity, citation-based strategies contend these would be articles of least trustworthiness.

3.2 Comparing Trust Algorithms

In the previous section we introduced different techniques for trust calculation. Now, we examine these methods comparatively. Our dimensions for comparison are not intended to be exhaustive. Instead, we choose attributes which highlight the strengths/weaknesses of each approach and reflect the design decisions commonly faced by collaborative trust systems. Table 3 summarizes the comparative merits of these algorithms.

3.2.1 User Identifier Persistence

For systems that include some notion of author reputation (which per Sec. 2.4 should be *all* systems), it is desirable that identifiers be persistent so one's contributions may be tracked throughout time. However, due to (1) anonymous editing, and (2) ill-intentioned users – this is not always the case.

Comparator	Persist	NLP	Meta	Cite	Sec.
Persistent IDs critical	Yes	No	Feature dependent	No	§ 3.2.1
Human involvement	Implicit Feedback	Corpus Building	Corpus Building	Implicit Feedback	§ 3.2.2
Integrates ext. data	No	n-grams	Yes	Yes	§ 3.2.3
Efficiency	Sufficient	Variable	Good	Sufficient	§ 3.2.4
Portability	See Table 4				§ 3.2.5

Table 3: Algorithm Comparison Summary

Wikipedia allows users to edit *anonymously*, whereby their IP addresses become used as identifiers. In such cases, it is unreliable to assume there is a 1:1 mapping between an IP address and an editor. A single public computer may have many users, and a single user may use computers in multiple locations. Further, a single computer may have a dynamic IP such that its addressing is not constant over time. Thus, it seems unreasonable to praise or punish IP identifiers for fear of collateral damage.

Even so, there exists a tension between anonymous and *registered* users (those with a persistent username/password). Nearly 80% of vandalism is committed by anonymous users [61], who contribute only 31% of all article edits [17]. Goldman [32] notes that anonymous users are sometimes treated as "second class citizens" and that their edits undergo additional scrutiny.

An obvious suggestion is to make all community members register, which is problematic for two reasons. First, Wikipedia (and its parent, the Wikimedia Foundation) is adamant in supporting anonymous editing, as it provides both convenience and privacy. Second, malicious users can still manipulate registered accounts to their benefit.

For example, one of the most common abuses leveraged at reputation systems is the Sybil attack [28]. New users must be given an initial trust value, and if the reputation of an account ever falls below that threshold, then it may be easier for an attacker to create a new account rather than repairing the reputation of the existing one. Wikipedia's barrier-to-entry – a CAPTCHA solve – seems ineffective in this regard since it has been shown that such protections can be evaded cheaply and at scale [47]. As a result,

trust systems must set initial values extremely low. Thus, new or casual users may be perceived just as anonymous users – "second-class" participants.

Choosing to be a registered editor does have benefits. Notably, the IP addresses associated with registered accounts are treated as private information[8], which may hamper some analysis. For example, the WikiScanner tool [16] detects conflicts-of-interest based on IP geo-location (*e.g.,* Edits from an IP from Redmond, Washington to the "Microsoft" article might warrant extra scrutiny). Similarly, [61] computes geographical reputations based on geo-location that prove effective in predicting the behavior of new users. Such analysis is not possible when IP addresses are not available.

So what do these issues mean for trust systems? Certainly, systems that arrive at user-reputations associatively (citation-based, NLP-based) are less affected than those that compute user reputations directly (content-driven, metadata-based). For the latter class, it is important that mechanisms are in place to evaluate users in the absence of history (for example, the spatial reputations of [61]). Secondly, if trust values are used to incentivize good behavior (see Sec. 4.2.4), then users will be rewarded for creating and maintaining persistent identifiers, lessening the severity of the issue.

3.2.2 Degree of Autonomy

We next examine the degree of *autonomy* at which each of the proposed techniques operates. That is, what role do humans play in the computation of trust values? We divide the space into three divisions: (1) *Corpus-driven*, (2) *Explicit-feedback*, and (3) *Implicit-feedback*.

Corpus-driven: First, we consider models which require no human intervention to evaluate a revision at the time it is committed. This includes NLP-based and metadata-driven strategies – precisely those which employ machine-learning and are corpus-driven. Whether knowingly or implicitly, humans have aided in labeling the corpora used to construct scoring models. Since models are pre-computed, they can be readily used to quantify revision quality. However, there are start-up costs associated with such approaches since corpora must be amassed for this purpose.

[8]Wikipedia does retain such data and makes it visible to a small set of extremely trusted users (`checkusers`). IP addresses are only investigated when it is suspected that abuse is being conducted via multiple accounts under the control of one individual.

Explicit-feedback: Second, are systems which require human involvement external of normal *wiki* actions in order to produce trust values. In our survey, we consider no systems of this type because they are uncommon, intrusive, prohibit automatic trust calculation, and have marginal cost. Nonetheless, such systems do exist in literature [42] and are in active use [15].

Such systems often manifest themselves as dialog boxes which allow a user to rate the quality of an article from an enumerated set of options. In other words, such systems collect *feedback*, subjective observations which form the basis for well-known reputation computations [36, 38].

Implicit-feedback: Most interesting are the content-driven and citation-based techniques which non-intrusively produce feedback by monitoring typical *wiki* behavior. For example, Adler's [19, 20] content-driven approach considers the removal of content to be an implicit *negative* feedback against that content and its authors. Citation-algorithms consider the citation of an article to be an implicit *positive* feedback about article quality.

Thus, these approaches can use well known feedback-aggregation strategies to produce behavior-predictive values. Beyond this, many systems have leveraged properties of collaborative environments to overcome complications typical of reputation management. For example, Adler's approach succeeds in holding feedback *providers* accountable – a challenge in traditional systems. Consider that an editor B who removes all the additions of A in an attempt to discredit him will be jeopardizing his own reputation, since if A's contribution is restored, it will be B who is punished. Similarly, B cannot simply praise the edits of A. Instead, B must actually edit the article, and then both the edits of A and B will be judged by subsequent editors. Further, since edit-magnitude is a factor, ballot-stuffing attacks are averted.

Similarly, many reputation systems are vulnerable to the "cold-start problem" (and thus, Sybil attacks, see Sec. 3.2.1) since multiple feedbacks may be required before meaningful values can be computed for an entity. West [61] overcomes this issue by broadening the entity under evaluation, leveraging the sociological property of homophily [46].

Implicit-feedback approaches are not without drawbacks, the most significant of which is *latency*. With content-persistence, multiple subsequent revisions are the best measure of a previous revision's quality. Thus, it may take considerable time for rarely edited articles to get their content vetted. Such latency could be significant in small communities where there are few feedback providers (see the 'intra-magnitude' portion of Sec. 3.2.5).

Latency is far worse for citation-based approaches. The decision to cite an article can speak to quality *only* when the citation was made. It is unreasonable to assume that the citation network evolves as dynamically as the underlying content (*i.e.,* such metrics are poor for vandalism detection).

Latency aside, the primary criticism of citation approaches is whether or not a citation actually constitutes a subjective feedback. That is, do *wiki* citations occur because individuals actually trust the page being cited, or is convention simply being followed? Wikipedia does specify linking conventions [14] which would skew the calculation of link-ratio and PageRank-like metrics. For example, the policy states one should "...link only the first occurrence of an item" on an article and that "...religions, languages, [and] common professions ..." should generally not be cited. Even the link-ratio authors recognize that proper nouns tend to be linked more frequently than well understood concepts (*e.g.,* love) [45]. These factors seriously challenge the extent to which citation-based metrics are measuring trust.

3.2.3 Integration of External Data

A *wiki* environment, in and of itself, provides a wealth of information which enables the calculation of trust values. However, several techniques distinguish themselves in that they are able to use data *external* to the wiki for on-wiki evaluation. The advantages of using external data are numerous. First, such data is outside the immediately modifiable realm, making it difficult for malicious users to manipulate. Additionally, smaller *wiki* installations may have sparse data, which external information could bolster.

Citation-based strategies can utilize external data by expanding the scope of their network graph. Rather than considering the internal hyperlink structure of the *wiki*, HITS/PageRank could measure incoming citations from outside the *wiki*. In other words, the algorithms would be used precisely as they are for search engine ranking – by crawling the entire Internet and processing citations. Then, the scores for articles could be interpreted comparatively. Indeed, an external citation of a *wiki* article seems to be a stronger endorsement of article trust than an internal one (per Sec. 3.2.2).

Only the most recent NLP-based works have utilized external data, in particular that of Wang [60] in their syntactic and semantic n-gram analysis. Whereas previous works pre-computed n-gram probabilities using a general corpus or the article itself as a topic-specific corpus – Wang uses the top-50 search engine results for an article title as the corpus for that article's

probabilities. Scalability issues aside, external data succeeds in increasing the breadth of such corpora. Further, one could imagine that web-corpora make n-grams more adaptable than other types. For instance, breaking news events may cause a revision to deviate from an article's typical scope. While typical corpora would cause such an addition to be viewed as unusual or deviant – Internet sources would likely have updated information and a constructive revision would be marked as such.

Finally, metadata approaches provide some of the richest opportunities for the use of external data. Indeed, the number of `JOINS` between metadata fields and external data seems limitless, although few have been investigated in literature. As an example, consider the IP address of an editor (a metadata field). In turn, that IP address could be used to: geo-locate the editor (and determine their local time-of-day or day-of-week), determine the editor's ISP, investigate the blacklist status of the IP address, or scan for open ports to determine if the IP is an open proxy.

The sheer size of feature-space available to researchers is undoubtedly one of the strongest assets of the metadata approach. However, critics may argue that metadata-feature are "a level removed" from what is really of interest – the content. Rather than encouraging contributors to author quality content, metadata-based features introduces a host of other variables into the evaluation process. Furthermore, there is the possibility of collateral damage and introducing disincentives to participation. Imagine a rule like "if an editor is from region x the trust in their edits should be reduced by y." Though it may be based on empirical evidence, such a rule may discourage innocent editors from the same region.

3.2.4 Computational Efficiency

Although theoretical advancements are useful, for a trust calculation system to actually be useful it needs operate efficiently at the *wiki* scale. Certainly, English Wikipedia suggests this may be computationally non-trivial. As of this writing, Wikipedia averages 1.5 edits/sec. in English, and 4 edits/sec. across all language editions [17] – and it is reasonable to assume peak loads may exceed these rates by an order of magnitude or more.

In the literature, we are aware of two works which cite concrete throughput figures. The NLP approach of Potthast [51] states it can handle 5 edits/sec., while the metadata technique of West [61] claims 100+ edits/sec.[9].

[9]Latency is not believed to be a significant issue. Although production systems make

While WikiTrust [18] (content-persistence) cites no explicit throughput numbers, its live implementation suggests it is capable of sufficient scalability. Similarly, Cluebot [1] speaks to the scalability of lexical NLP techniques.

Thus, significant scalability questions remain about (1) citation-based and (2) predictive NLP (*i.e.,* n-grams), and we examine each in turn.

It would seem that no matter the citation-based algorithm, a considerable amount of pre-processing would be required to construct the initial network graph. However, once this is complete, the link-ratio algorithm of McGuinness could trivially update values in an incremental fashion (as each index term has a value independent of all others). Probability-based citation algorithms like PageRank/HITS are more complex, given that an evolving network structure could alter probabilities for a large number of nodes. Nonetheless, incremental update techniques have been developed for PageRank, and the Wikipedia network is orders of magnitude smaller than the Internet-scale space these algorithms were designed to process. Further, since citation-based trust is ineffective for short-term assessments (*e.g.,* vandalism), some delay in trust value calculation is acceptable.

Predictive NLP techniques also require a large amount of pre-processing to have n-gram probabilities ready to compare against new ones in an edit `diff`. The distinguishing factor is *when* this pre-processing can be performed. If one uses a large and general-purpose corpus, there is little issue in having probabilities readily available at edit-time. However, research has shown that domain-specific probabilities are advantageous. This means, at a minimum (supposing the previous version of the article is treated as a corpus), probabilities would need re-calculated for each article after every edit. In the worst case are dynamic web-based corpora like those proposed by [60], who used the top-50 web results for an article's title to be the training corpus. Such a massive amount of text-processing (and the considerable bandwidth costs) seems problematic at scale.

3.2.5 Technique Portability

Though our analysis herein is focused on the English Wikipedia, it is important to realize there are many *wiki* installations across the Internet. For instance, Wikipedia has 273 language editions and nearly a dozen sister

API calls [11] to Wikipedia, adding latency, such approaches could conceivably run on the Wikimedia servers if they were deemed sufficiently important.

Approach		Language	Standard	Magnitude
Content-persist		✓	✓	
NLP	Lexical			✓
	n-gram	✓		✓
Metadata-based		✓		✓
Citation-based		✓	✓	

Table 4: Portability of Trust Approaches

projects (and their language editions). Additionally, `wikia.com` – a centralized *wiki* hosting service – supports over 100,000 *wikis* [9]. These examples likely constitute only a trivial fraction of installations on the Internet. It is likely that most of these communities lack the tools, vigilance, and massive user-base that enables English Wikipedia to thrive.

Thus, automatic calculation of trust values seem especially useful in such installations. We consider three dimensions of portability for our trust techniques: (1) *intra-language* (*e.g.,* as English Wikipedia relates to French Wikipedia), (2) *intra-purpose* (*e.g.,* as Wikipedia relates to Encyclopædia Dramatica), and (3) *intra-magnitude* (*e.g.,* as Wikipedia relates to a small-scale installation). Table 4 indicates which algorithms can be transitioned between dimensions with no/trivial modification to their approach.

Intra-language: First, we address the portability of techniques across different natural languages. Intuitively, such a transition is most problematic for NLP-based measurement, but to a surprisingly small extent. Lexical techniques (*e.g.,* bad-word regexps) would need to be localized, but semantic and syntactic measures (*e.g.,* n-gram probabilities) can be used so long as they are calibrated over corpora in the language of interest. Meanwhile, content-persistence techniques require only that the natural language be delimited in some way (presumably at word or sentence granularity). It is reasonable to assume most natural languages have this characteristic.

Intra-purpose: Second is the issue of intra-purpose portability. Are trust mechanisms tuned for Wikipedia's encyclopedic expectations, or do these expectations hold for content in general? Both NLP and metadata-based approaches seem challenged by such a transition. The biggest unknown for NLP is how predictive measure (*i.e.,* n-grams) might operate when *novel*

content is being generated (*e.g.,* imagine collaboratively authoring a fiction novel), rather than summarizing some existing body of knowledge (as with an encyclopedia). Similarly, metadata-based IQ metrics would also be sensitive to change, as they were manually crafted for encyclopedic use by [58] (though versions do exist for generalized web documents [64]).

Intra-magnitude: Finally, we consider the magnitude of the *wiki* under investigation and in particular how smaller wikis might affect trust computation. Content-persistence methods are dependent on the implicit feedback made by subsequent editors to an article. Such assessments may be considerably latent in a *wiki* with a low edit volume. Citation-driven approaches could also be considerably hampered. Consider that a *wiki* with few editors is unlikely to generate much content, and in turn, the citation graph is likely to be small and sparse. Such graphs are not ideal for calculating link-ratios or internal PageRank/HITS scores.

3.2.6 Comparative Effectiveness

Perhaps the most obvious question regarding the varied techniques is, *"which works best?"* – and unsurprisingly, a definitive answer is not possible.

For systems that compute article-granularity trust, the standard approach has been to examine articles that were specially tagged by the editing community (*e.g.,* "Featured Article" or "Article Needs Cleanup"). Though this may provide some basis for comparison, it is unclear the completeness or accuracy of these taggings (less than 0.1% of articles are 'featured').

More satisfying is the recent vandalism corpus and subsequent detection competition of Potthast *et al.* [52]. The corpus is composed of 32,000 revisions, labeled by crowd-sourced annotators. For the detection competition (which withheld labels for half the corpus), 9 different schemes were submitted, encompassing 55 different features, all of which are discussed in the competition summary [52].

Three of our methodologies, (1) content-driven, (2) NLP-based, and (3) metadata-based were well represented in the competition (only citation-based is not, which does not apply well at revision-granularity). An NLP approach based on [51] won the competition, with WikiTrust [18] (content-persistence) finishing second. We believe these results should be interpreted cautiously, as each system is only a single, non-comprehensive, point of reference into a domain. Further, the competition only gauged how well systems apply in

the domain of vandalism detection and not across the entire trust spectrum.

Most importantly, [52] reports that a meta-classifier built from all competition entries significantly outperforms the single winning classifier. Thus, differing strategies capture unique sets of vandalism, validating that trust is an issue best approached via multiple methodologies.

4 Usage of Trust Values

While the calculation of accurate trust values is important, these values are only useful if they are effectively communicated to the end-user or utilized internally in a way that benefits the end-user. Thus, in this section we survey proposals and implemented systems that present/utilize trust values and discuss some of the challenges in developing systems of this kind.

4.1 Interpreting Trust Values

One of the biggest challenges in utilizing trust values is that they must be *relatively interpreted*. None of the systems we have surveyed are capable of computing values that can be read in an absolute capacity. As a result, no definitive statements can be made about what is 'good' and 'bad' and comparative analysis becomes necessary.

Comparative values are not ideal. Unlike in search-engine retrieval, it seems unlikely that a *wiki* user would need to determine which of two documents is most trustworthy. It is more likely that they would wish to know the trust of an article in isolation.

Two strategies attempt to impart meaning onto values: (1) Treating values as a classification problem and applying thresholds based on empirical evidence, and (2) Normalizing values to make them presentation-friendly. The first approach, as discussed in Sec. 2.3.2, requires training corpora to be amassed. While simple to build for certain subsets of the trust spectrum (*i.e.,* vandalism), this is a difficult approach for more fine-grained analysis. Further, thresholds are often drawn based on a tolerance for false-positives, not the need for absolute accuracy.

The second approach, normalization, is often used for presentation purposes. For example, trust values on the range $[0, 1]$ are more human-friendly than raw values. Of course, normalized values are arbitrary and perhaps even

Approach	Granular.	Tasks (Sec. 4.2)
Content-persist	Fragment, Author	Fragment trust, Revision selection, User privileges
NLP	Revision	Anti-vandalism
Metadata-based	Article, Revision	Article trust, Anti-vandalism
Citation-based	Article	Article trust

Table 5: Describing the "preferred granularity" of each approach – and the tasks which computed values are most useful at optimizing.

deceptive to human users. For example, articles on a poor quality *wiki* could have full normalized trust because they are the "best among the worst."

Alternatively, one can simply embrace the relative nature of trust values and ignore mapping attempts. Such is the approach of *intelligent routing systems*, such as [61], which we discuss further in Sec. 4.2.2.

4.2 Application of Trust Values

In this section, we examine the application of trust values to different tasks on Wikipedia. For each task, we first describe how the Wikipedia community currently performs the task (*i.e.,* the status quo). Then, we demonstrate how the application of trust values may optimize that task, making it more efficient, accurate, or intuitive. Table 5 summarizes the approaches which excel at each task (often due to a preference for calculating trust at a specific granularity). Our choice of tasks is not intended to be comprehensive, but reflect some of the most prominent proposals in the literature.

The work of Goldman [32] suggests the need for such optimizations is pertinent – as Wikipedia's dwindling editing force struggles to maintain order as readership and mischief continue to expand.

4.2.1 Visual Display of Trust

Status Quo: Perhaps the most straightforward use of trust values is to present them directly to the end-user, adjacent to the article or text fragments they describe. The Wikipedia software has no functionality for this purpose at present – though it has been a popular proposal among researchers.

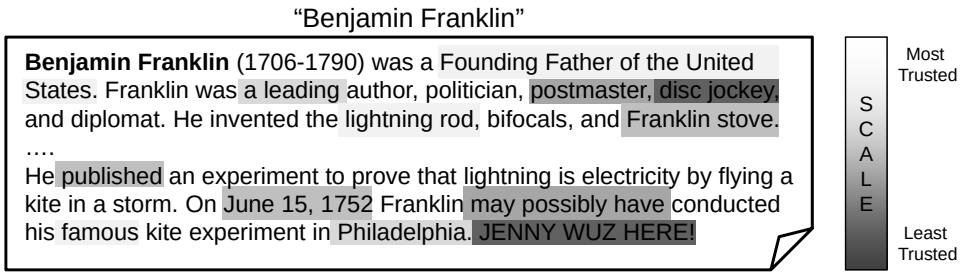

Figure 7: Example using text-coloring to display trust

Trust Application: Several authors [19, 26, 45] propose the colorization of fragment text as an intuitive and non-intrusive way to present trust values. A live browser plug-in utilizing the technique has been developed [18]. Figure 7 displays an example of the proposed output.

More simply, a suggestion has been to simply display the numerical trust value of the article, on the article itself[10] [42]. Of course, public exposure of trust values can lead to problems with interpretation (Sec. 4.1) or encourage evasion (Sec. 4.3) and may be a reason such proposals are yet to gain traction.

4.2.2 Damage Detection

Status Quo: Broadly, three strategies are currently currently used to detect and undo damaging edits (*i.e.,* vandalism). First, is the automatic reversion of poor edits by autonomous *bots* – of which the NLP-based ClueBot [1] would be the characteristic example. Second, is the use of software assistants to present edits to human users and asks them to make determinations. Huggle [6] is the most popular example, which prioritizes edit display using a simple and manually-authored rule set (*e.g.,* show anonymous-user edits before those of registered ones). Finally, there are damaged edits discovered purely by human chance or brute-force. For example, editors often monitor changes to articles they are interested in via customized *watchlists*, or do brute-force patrol by watching the "recent changes" feed.

[10]While a valid proposal, we note that the cited system relies on explicit user-provided feedback and is not capable of automatic trust calculation. Thus, the display of these numerical values side-steps earlier issues involving relative interpretation of trust values.

Trust Application: We first address the creation of smarter Wikipedia bots. Bots are attractive since they act quickly and at zero marginal cost. However, community standards are such that there is minimal tolerance for false-positives by such bots. Thus, in the current state-of-the-art such bots can only address the most "low hanging fruit." The comparison of multiple detectors by Potthast [52] showed that only one system was capable of nearly false-positive free performance (at any level), and it was only capable finding 20% of damage at such high accuracy.

Given this, we believe software-assisted human detection should be a point of focus. Relative trust values can be well leveraged to build *intelligent routing tools* [25], which direct humans to where their efforts are most needed (*i.e.,* probable damage). At present, this technique is best leveraged by the STiki tool [61], which has a shared priority queue of revisions.

4.2.3 Revision Selection

Status Quo: While vandalism detection focuses on determining if the last edit to an article was damaging, *revision selection* tackles the more general problem of determining which version in an article's history is 'best.' The selected version can then be the default displayed under certain criteria or used to build trusted snapshots for other purposes.

On Wikipedia, such functionality is leveraged by a software extension called `FlaggedRevs` [4]. One use-case of the extension – "Pending Changes" – is currently active on several foreign language editions and under trial on the English Wikipedia [24]. The system prevents the revisions of anonymous editors from being publically displayed (on certain protected pages) until they have been approved by a trusted community member (a `reviewer`).

Trust Application: For Pending Changes is concerned, trust values could be used to reduce reviewer workload by not requiring approval for highly trusted revisions. However, more interesting than its anti-vandalism use is how `FlaggedRevs` might be used to flag trusted revisions occurring many revisions in the past and on articles without explicit review.

For example, projects are creating static snapshots of Wikipedia for use by schools and for DVD/print distribution. Clearly, it is desirable that such snapshots contain the 'best' versions of an article possible – and that part of that definition should include 'damage-free.' Content-persistence trust is well suited for this task since it can evaluate revisions using the benefit of

(a)

(b)

Figure 8: User reputation; (a) warnings, and (b) barnstars

hindsight. However, 'currency' is also like to be a factor in what defines the 'best' revision. The most recent edits – those which likely include the most current information – are precisely those which we know the least about under content-persistence. Metadata or NLP techniques could prove helpful in this regard, but how to best weigh these factors remains an open research question. Regardless, any automation of the process is likely to be an improvement over the manual inspection currently employed as a safe-guard.

4.2.4 User Privileges

Status Quo: Editing *privileges* on Wikipedia include not just the advanced permissions delegated to trusted participants, but also the privilege to simply edit the encyclopedia which is sometimes revoked from troublesome users.

Wikipedia has a semi-formal mechanism by which users can lose reputation and privileges. Editors committing damaging edits will be communicated increasingly stern warnings (see Figure 8a), which if ignored, will eventually lead to blocks/bans [30]. Individual accounts, single IP addresses, and IP ranges can be blocked/banned as needed to stop abuse.

In contrast, there is little formality in the way reputation is amassed. While prolific editors may advance to `administrator` status and have extensive personal interaction histories, the vast majority of editors likely reside in a vast gray area where informal measures dominate. For example, *edit count* is sometimes viewed as a measure of reputation, though [29] observes this to be a poor measure. Further, *barnstars* – personalized digital tokens of appreciation (see Figure 8b) – are sometimes awarded between users [41].

Trust Application: As Adler *et al.* [20] note, the integration of user-level reputations into a *wiki* setting is important because it can *incentivize* constructive behavior. Unfortunately, Wikipedia has seemed to take the opposite approach by simply punishing miscreants.

Wikipedia has long championed the open-editing model, with minimal hierarchy among contributors and few restrictions. However, Goldman [32] notes that Wikipedia's labor shortage may force new built-in protections (*e.g.,* locking articles, pending changes, *etc.*) to mitigate poor behavior. With these protections comes the need for *new permissions* to manage them (or be exempt from them) will be inevitable. User trust could provide a means to automate the delegation and revocation of such rights, while providing a degree of robustness[11].

4.3 Cautions for Value Usage

Though the application of trust values in *wiki* settings is primarily viewed a a benefit, we briefly discuss the potential drawbacks of integrating trust values into collaborative software. These drawbacks are not intended to discourage the use of collaborative trust, but rather to highlight some design decisions about which developers should be cautious.

First, automatic tools and prioritization mechanisms may lead to a false sense of security and over-confidence. For example, if the STiki [8] anti-vandalism tool poorly classifies an edit, it will receive low priority, and may never be reviewed by a human. Tools like STiki and Huggle [6] have reduced the numbers of editors doing brute-force vandalism patrol, though the affect this has on anti-vandalism efforts is unknown.

Second, the exposure of trust values may provide malicious users insight into how trust values are calculated, permitting evasion. The most prominent example of this is Wikipedia's Edit Filter [2], which uses a manually generated rule set and can prevent edits from being committed. If an edit is disallowed, the reader will be informed of such – encouraging them to reshape their edit into something slightly more constructive (or evasive). Thus, profanity may be obfuscated to evade the filter. Not only will this evade the Edit Filter, but it may also evade downstream mechanisms (*e.g.,* bots) which

[11] Wikipedia has a psuedo-permission called `autoconfirmed`, to which registered users *automatically* advance after 10 edits and 4 days (post-creation). `Autconfirmed` users need not solve CAPTCHAs and have other minor benefits. Clearly, given the ease of manipulating a metric like "edit count", this could be a vector for abuse.

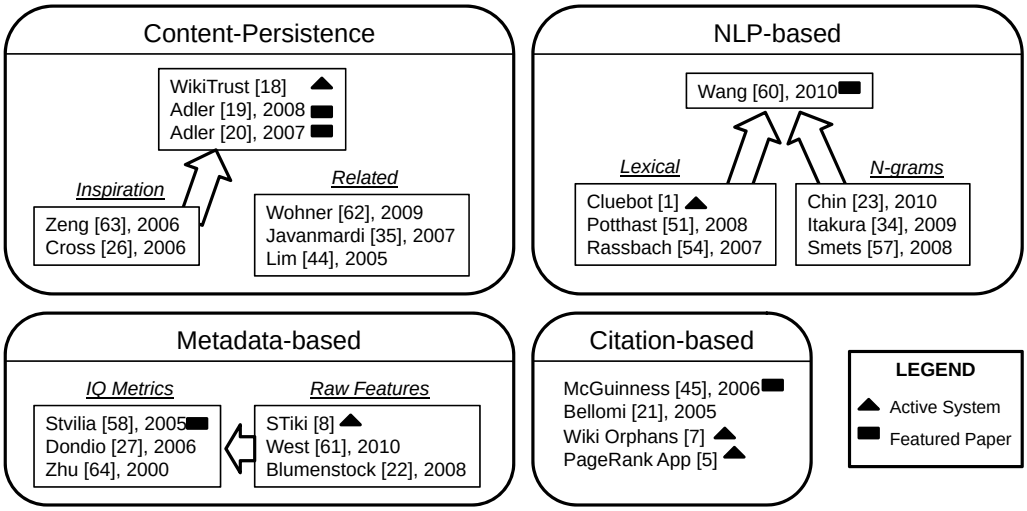

Figure 9: Relationship Among Publications (Wikipedia-centric)

could have caught the original edit. Fortunately, those who damage articles seem poorly motivated. Priedhorsky *et al.* [53] observes that 71% of damaging edits exhibit 'nonsense' or 'offensive' attributes. However, [32] indicates that Wikipedia's growing popularity will invite motivated malicious users, such as spammers, who have financial incentive to evade protections.

Finally, the exposure of user-granularity trust presents a unique set of challenges. Adler [20] advocates the display of user reputation values, arguing that public values will incentivize users to behave well. Nonetheless, there are counter-arguments. Wikipedia encourages an open-editing model where everyone is free to edit the work of others. User trust values could create a fine-grained hierarchy of editors which would create a barrier-to-entry and less democratic collaboration. Public reputation may also lead editors to over-emphasize the importance of their own reputations. This may lead to editors doing solely what is best for *themselves* as opposed to the *encyclopedia*. For example, under content-persistence trust editors may avoid editing breaking news topics, as their contributions are likely to be undone as the story evolves (regardless of their accuracy at the time of editing).

5 Conclusions

Herein, we have surveyed four different approaches to calculating trust for collaborative content and discussed how these trust values can benefit the cooperative process. As Figure 9 shows, these works are supported by a large body of prior literature and related research. Each proposal has its relative merits and has been shown successful via evaluation, yet there is evidence that the state-of-the-art still has many challenging, open research questions.

Though it is evident these systems are computing meaningful values (per their performance), it is not always clear to what extent these values speak to the actual *trust* one should place in an article/user/fragment. Of course, this is complicated by the many definitions of trust in literature and the fact that few of them make for easy quantification. To side-step this issue, most authors focus on the most trivially untrustworthy of edits (*i.e.,* vandalism) to gain traction on the problem. It remains to be seen if these vandalism-centric values are capable of meaningfully quantifying contributions across the entirety of the trust spectrum.

Perhaps the most encouraging aspect of the different approaches is that they capture *unique* kinds of poor behaviors. As a recent vandalism-detection competition showed, meta-detectors built from different approaches significantly outperform any individual system. Thus, understanding how these approaches can interact to produce higher-order classifications is an important step moving forward.

Moving forward will also involve study of *wiki* environments other than Wikipedia. While Wikipedia is a large entity with available data, its community dynamics may be far different than those elsewhere online. Understanding how trust systems can work in generic collaborative environments is important to their application elsewhere. Further, most *wikis* rely on text-based content. How techniques might adapt to collaborative systems based on images or data will be an interesting evolution.

Regardless, the potential for trust systems in collaborative systems is large. For established systems like Wikipedia, they may ease maintenance concerns and allow editors to focus on content development. For emerging systems, trust can allow the community to measure its progress and highlight content which may best serve readers. On the whole, preventing readers from mis-information is crucial as society becomes increasingly reliant on collaborative knowledge.

References

[1] ClueBot. http://en.wikipedia.org/wiki/User:ClueBot.

[2] Edit filter. http://en.wikipedia.org/wiki/WP:Edit_filter.

[3] Encyclopædia Dramatica. http://www.encyclopediadramatica.com.

[4] FlaggedRevs. http://www.mediawiki.org/Extension:FlaggedRevs.

[5] Google toolbar. http://www.google.com/toolbar/.

[6] Huggle. http://en.wikipedia.org/wiki/Wikipedia:Huggle.

[7] Orphaned articles. http://en.wikipedia.org/wiki/CAT:ORPH.

[8] STiki. http://en.wikipedia.org/wiki/Wikipedia:STiki.

[9] Wikia. http://www.wikia.com/.

[10] Wikipedia. http://www.wikipedia.org.

[11] Wikipedia API. http://en.wikipedia.org/w/api.php.

[12] Wikipedia: Featured article criteria. http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria.

[13] Wikipedia glossary. http://en.wikipedia.org/wiki/WP:Glossary.

[14] Wikipedia manual of style. http://en.wikipedia.org/wiki/WP:LINK.

[15] Wikipedia's public policy initiative. http://outreach.wikimedia.org/wiki/Public_Policy_Initiative.

[16] WikiScanner. http://wikiscanner.virgil.gr/.

[17] Wikistats: Wikimedia statistics. http://stats.wikimedia.org/.

[18] WikiTrust. http://wikitrust.soe.ucsc.edu/.

[19] B. Adler, J. Benerou, K. Chatterjee, L. de Alfaro, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *WikiSym 2008: International Symposium on Wikis*, 2008.

[20] B. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *WWW 2007, Proceedings of the 16th International World Wide Web Conference*, 2007.

[21] F. Bellomi and R. Bonato. Network analysis for wikipedia. In *WikiMania '05: The First International Wikimedia Conference*, 2005.

[22] J. E. Blumenstock. Size matters: Word count of a measure of quality on Wikipedia. In *WWW '08: Proc. of the 17th Intl. Conference on the World Wide Web*, pages 1095–1096, 2008. (Poster paper).

[23] S.-C. Chin, W. N. Streeta, P. Srinivasan, and D. Eichmann. Detecting Wikipedia vandalism with active learning and statistical language models. In *WICOW'10: Workshop on Info. Credibility on the Web*, 2010.

[24] N. Cohen. Wikipedia to limit changes to articles on people. *New York Times*, page B1, August 25, 2009.

[25] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing*, pages 1037–1046, 2006.

[26] T. Cross. Puppy smoothies: Improving the reliability of open, collaborative wikis. *First Monday*, 11(9), September 2006.

[27] P. Dondio, S. Barrett, S. Weber, and J. M. Seigneur. Extracting trust from domain analysis: A case study on the Wikipedia project. In *Autonomic and Trusted Computing*, volume 4158, pages 362–373. Springer Berlin/Heidelberg, 2006.

[28] J. Douceur. The Sybil attack. In *1st IPTPS*, March 2002.

[29] P. K.-F. Fong and R. P. Biuk-Aghai. What did they do? Deriving high-level edit histories in wikis. In *WikiSym '10: Proceedings of the Sixth International Symposium on Wikis and Open Collaboration*, July 2010.

[30] R. S. Geiger and D. Ribes. The work of sustaining order in Wikipedia: The banning of a vandal. In *CSCW '10: Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 117–126, 2010.

[31] J. Giles. Internet encyclopedias go head to head. *Nature*, 438:900–901, December 2005.

[32] E. Goldman. Wikipedia's labor squeeze and its consequences. *Journal of Telecommunications and High Technology Law*, 8, 2009.

[33] S. Hao, N. A. Syed, N. Feamster, A. G. Gray, and S. Krasser. Detecting spammers with SNARE: Spatio-temporal network-level automated reputation engine. In *18th USENIX Security Symposium*, 2009.

[34] K. Y. Itakura and C. L. Clarke. Using dynamic Markov compression to detect vandalism in the Wikipedia. In *SIGIR '09: Proceedings of the 32nd International ACM SIGIR conference on Research and Development in Information Retrieval*, 2009. (Poster paper).

[35] S. Javanmardi and C. V. Lopes. Modeling trust in collaborative information systems. In *Proceedings of Collaborative Computing: Networking, Applications, and Worksharing*, pages 299–302, 2007.

[36] A. Jøsang, R. Hayward, and S. Pope. Trust network analysis with subjective logic. In *Proc. of the Australasian Comp. Science Conf.*, 2006.

[37] A. Jøsang, C. Keser, and T. Dimitrakos. Can we manage trust? In P. Herrmann, V. Issarny, and S. Shiu, editors, *Trust Management*, volume 3477 of *Lecture Notes in Computer Science*, pages 93–107. 2005.

[38] S. D. Kamvar, M. T. Schlosser, and H. Garcia-molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proceedings*

of the Twelfth International World Wide Web Conference, 2003.

[39] D. Kaplan. Hackers use German Wikipedia to spread malware. `http://www.scmagazineuk.com/hackers-use-wikipedia-article-to-spread-malware/article/106455`.

[40] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[41] T. Kriplean, I. Beschastnikh, and D. W. McDonald. Articulations of wikiwork: Uncovering valued work in Wikipedia through barnstars. In *CSCW '08: Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work*, pages 47–56, 2008.

[42] T. Lefévre, C. D. Jensen, and T. R. Korsgaard. WRS: The Wikipedia recommender system. In *IFIPTM '09: Proceedings of Trust Management III*, pages 298–301, West Lafayette, Indiana, USA, 2009.

[43] B. Leuf and W. Cunningham. *The Wiki Way: Quick Collaboration on the Web*. Addison-Wesley, 2001.

[44] E.-P. Lim, B.-Q. Vuong, H. W. Lauw, and A. Sun. Measuring qualities of articles contributed by online communities. In *Proc. of the ACM/WIC/ACM Intl. Conf. on Web Intelligence*, pages 81–87, 2006.

[45] D. L. McGuinness, H. Zeng, P. D. Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigation into trust for collaborative information repositories: A Wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*, 2006.

[46] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 2001.

[47] M. Motoyama, K. Levchenko, C. Kanich, D. McCoy, G. M. Voekler, and S. Savage. Re: CAPTCHAs - Understanding CAPTCHA-solving services in an economic context. In *USENIX Security '10: Proceedings of the 19th USENIX Security Symposium*, August 2010.

[48] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford, 1999.

[49] B. Pershing. Kennedy, Byrd the latest victims of Wikipedia errors. `http://voices.washingtonpost.com/capitol-briefing/2009/01/kennedy_the_latest_victim_of_w.html`.

[50] S. Pogatchnik. Student hoaxes world's media on Wikipedia. `http://www.msnbc.msn.com/id/30699302/`.

[51] M. Potthast, B. Stein, and R. Gerling. Automatic vandalism detection in Wikipedia. In *Advances in Information Retrieval*, 2008.

[52] M. Potthast, B. Stein, and T. Holfeld. Overview of the 1st intl. competi-

tion on Wikipedia vandalism detection. In M. Braschler and D. Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops*, 2010.

[53] R. Priedhorsky, J. Chen, S. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *GROUP '07: Proc. of the 2007 ACM Conference on Supporting Group Work*, pages 259–268, 2007.

[54] L. Rassbach, T. Pincock, and B. Mingus. Exploring the feasability of automatically rating online article quality.

[55] M. Sahamia, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk email. In *Proceedings of the AAAI Workshop on Learning for Text Categorization*, 1998.

[56] J. Seigenthaler. A false Wikipedia 'biography'. `http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm`.

[57] K. Smets, B. Goethals, and B. Verdonk. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *Proc. of the AAAI Workshop on Wikipedia and Artificial Intelligence*, 2008.

[58] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proc. of the Intl. Conference on Information Quality*, pages 442–454, 2005.

[59] R. Wang and D. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Info. Systems*, 12(4), 1996.

[60] W. Y. Wang and K. McKeown. "Got you!": Automatic vandalism detection in Wikipedia with web-based shallow syntactic-semantic modeling. In *COLING' 10: Proceedings of the 23rd International Conference on Computational Linguistics*, 2010.

[61] A. G. West, S. Kannan, and I. Lee. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In *EUROSEC '10: Proceedings of the Third European Workshop on System Security*, 2010.

[62] T. Wöhner and R. Peters. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proc. of WikiSym '09*, 2009.

[63] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. In *International Conference on Privacy, Security, and Trust*, 2006.

[64] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *SIGIR '10: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 288–295, 2000.