

Date of publication XXXX 00, 2024, date of current version May 11, 2024.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

# Exploring Convergence in Relation using Association Rules Mining: A Case Study in Collaborative Knowledge Production

Jiahe Ling<sup>1</sup> and Corey B. Jackson<sup>1</sup>

<sup>1</sup>University of Wisconsin-Madison, Madison, WI 53706 USA

Corresponding author: Jiahe Ling (e-mail: jling9@wisc.edu).

**ABSTRACT** This study delves into the pivotal role played by non-experts in knowledge production on open collaboration platforms, with a particular focus on the intricate process of tag development that culminates in the proposal of new glitch classes. Leveraging the power of Association Rule Mining (ARM), this research endeavors to unravel the underlying dynamics of collaboration among citizen scientists. By meticulously quantifying tag associations and scrutinizing their temporal dynamics, the study provides a comprehensive and nuanced understanding of how non-experts collaborate to generate valuable scientific insights. Furthermore, this investigation extends its purview to examine the phenomenon of ideological convergence within online citizen science knowledge production. To accomplish this, a novel measurement algorithm, based on the Mann-Kendall Trend Test, is introduced. This innovative approach sheds illuminating light on the dynamics of collaborative knowledge production, revealing both the vast opportunities and daunting challenges inherent in leveraging non-expert contributions for scientific research endeavors. Notably, the study uncovers a robust pattern of convergence in ideology, employing both the newly proposed convergence testing method and the traditional approach based on the stationarity of time series data. This groundbreaking discovery holds significant implications for understanding the dynamics of online citizen science communities and underscores the crucial role played by non-experts in shaping the scientific landscape of the digital age. Ultimately, this study contributes significantly to our understanding of online citizen science communities, highlighting their potential to harness collective intelligence for tackling complex scientific tasks and enriching our comprehension of collaborative knowledge production processes in the digital age.

**INDEX TERMS** Association Rules, Citizen Science, Knowledge Production, Open Collaboration, Social Computing, Time Series Analysis

## I. INTRODUCTION

Knowledge production, the process of generating knowledge, essential in open collaboration platforms, involves generating, organizing, and disseminating information to enhance existing knowledge. These platforms, such as Wikipedia and open-source software development communities, facilitate collective intelligence by allowing individuals to contribute expertise and insights for communal benefit. A growing focus lies in online citizen science, engaging non-professionals in scientific research, exemplified by projects like Zooniverse's Gravity Spy. Here, volunteers analyze data collectively, aiding scientific discovery in fields like astronomy. This participatory model extends beyond content discovery, shaping scientific

understanding. However, as tasks grow more intricate, empirical evidence is vital to justify involving amateur contributors in coordination and communication-intensive projects.

Association rules analysis is a data mining technique utilized to discern relationships or associations between variables within extensive datasets. Its objective is to unveil intriguing patterns or correlations among items, predicated on their co-occurrence in transactions or events. Within social computing, association rules mining is harnessed to uncover patterns and relationships within social interactions and user-generated content, providing valuable insights into user behavior, community dynamics, and content engagement.

Previous studies have utilized clusters, tag clouds, and association rules analysis to investigate relationships among tags in collaborative tagging systems. However, there has been limited research on continuously examining how these relationships evolve over time.

This paper delves into the collaborative tagging practices within the Gravity Spy project to understand how knowledge is created and shared among non-experts. By analyzing association rules weekly, the study tracks how these metrics evolve over time. This approach enables the examination of the relationship between user-provided tags. Crucially, this research proposes a methodology to ascertain whether the community converges towards one or multiple hashtags for each image over time by identifying the point at which the series approaches a specific value through continuous reduction of variation.

This paper endeavors to investigate the phenomenon of ideological convergence in online citizen science knowledge production by introducing a novel measurement algorithm. It seeks to address the question: “How prevalent is ideological convergence in knowledge production?” Additionally, it aims to compare this new algorithm with traditional methodologies for analyzing convergence.

## II. LITERATURE REVIEW

The process of knowledge production involves generating, organizing, and sharing information to enhance existing knowledge. This process encompasses collecting data, gaining insights, and communicating these insights to others [1]. Knowledge can be explicit (formal, documented) or tacit (personal, experiential). Historically, knowledge production was centralized but has shifted towards decentralization with the advent of the Internet and information technologies [2], [3]. Decentralized production, exemplified by platforms like Wikipedia and OpenStreetMap, fosters inclusivity and diversity, leading to a richer knowledge ecosystem. However, challenges such as quality control and sustainability persist [4], [5], [6]. Solutions proposed include establishing common ground, leveraging user engagement, and infrastructure development. Collaborative tagging, as seen in citizen science projects like Galaxy Zoo and BeeWatch, exemplifies bottom-up categorization and information retrieval, though concerns about content quality remain. Collaborative tagging involves users applying descriptive tags to content, reflecting diverse viewpoints and interpretations, and collectively creating content organization.

### A. FOLKSONOMIES

Folksonomies, a collaborative tagging system utilized to categorize and organize digital content, are generated by a community of users who assign tags based on personal understanding and perspectives [7]. Observations by Makani

and Spiteri indicate a decline in unique tags over time within a knowledge management community, suggesting the establishment of a stable and domain-specific vocabulary [8]. Additionally, Kopeinik et al. demonstrated that tag curation significantly contributes to semantic stabilization in group learning projects [9]. Santos-Neto et al. proposed utilizing tagging activity distribution and user interest similarity metrics to enhance navigability in expanding knowledge spaces [10]. However, folksonomies embody a decentralized approach to content categorization, leading to inconsistencies, redundancies, and ambiguities in tag interpretation. For instance, synonymous spelling variants or acronyms pose challenges in observing knowledge production through tagging.

### B. ASSOCIATION RULE MINING (ARM)

Association rule mining (ARM) is a widely used technique for uncovering significant patterns in large datasets [11]. Folksonomies, prevalent collaborative tagging systems in social media, serve as valuable data sources for ARM applications. One key use of ARM in folksonomies is exploring patterns and relationships among user-generated tags, revealing insights into folksonomic dynamics, community formation, and information diffusion. Researchers have systematically investigated different projections of folksonomy structures to uncover meaningful tag associations [12]. By applying ARM to extensive folksonomy datasets, researchers have demonstrated the utility of discovered rules for various purposes, including tag recommendation, user profiling, resource organization, and community detection [13]. However, literature suggests that the utilization of ARM in folksonomies remains limited compared to other analytical approaches like tag clouds and tag-based recommendations. Further in-depth analysis of ARM within folksonomies is thus warranted to fully leverage this data mining technique in social media contexts.

### C. CONVERGENCE AND STATIONARITY

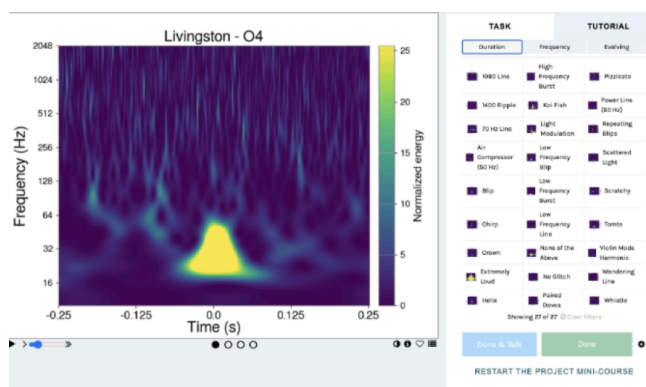
The term “convergence” carries varied meanings across different contexts. In economics, it typically denotes the diminishing disparities or distinctions among regions, entities, or variables over time [14]. In some other fields, the “convergence” of a time series is also used interchangeably with the term “stationarity,” which entails the constancy of statistical properties over time [15]. However, in this paper, convergence is defined more strictly: it signifies a time series approaching a specific value through continuous reduction of variation as time advances.

In collaborative learning, convergence is not determined through statistical testing, but rather conceptualized as an ideological alignment among individuals. Experiments and surveys are typically conducted to discern the factors influencing convergence and to ascertain whether individuals indeed reach agreement [16].

In summary, no quantitative (statistical) methodology currently exists to objectively measure the convergence of individuals towards a particular idea in collaborative learning. Existing methods can only detect if a time series diminishes disparities or achieves stationarity relative to another. This underscores the significance of this research, which aims to quantify the learning patterns of a collaborative tagging system using ARM as a time series. Subsequently, this paper proposes a methodology to identify the point in time at which the series approaches a specific value through continuous reduction of variation as time progresses.

### III. DATA

The research focuses on Gravity Spy, an online citizen science project where volunteers collaborate with researchers to identify and categorize noise signals in gravitational wave data. Volunteers assist by classifying datasets into known categories and developing new ones. They use a classification interface equipped with a Field Guide for reference and metadata options. As volunteers gain expertise, they progress to more advanced workflows. Additionally, volunteers engage in discussions on the Gravity Spy Talk platform, particularly regarding unidentified glitches which may signify new categories. Volunteers can propose new glitch classes by submitting detailed proposals, which are then evaluated by LIGO scientists. These proposals include images, descriptions, and proposed names for the new glitch classes. The process aims to formalize the creation of new glitch categories based on volunteer input. An example Gravity Spy classification interface is shown in **Figure 1**.



**FIGURE 1.** The Gravity Spy classification interface: an image of a glitch on the left, and the list of classification options on the right.

This study gathered digital trace data encompassing all comments posted by volunteers on Zooniverse between October 2016 and July 2023. Emphasis was placed on hashtags as they are the primary vehicles for linking image objects across the platform. Each entry in the hashtag dataset included a timestamp indicating the posting time, along with metadata such as board\_id, discussion\_id, the hashtag itself, and the username of the volunteer who posted the comment. To provide context, it is important to note that the platform

consists of various discussion boards, each containing multiple discussions. Within each discussion, numerous hashtags could be proposed. To streamline the analysis, comments posted on boards titled “Notes” were included. Unlike other discussion boards, “Notes” boards are associated with image subjects, suggesting that hashtags on these boards likely relate to new glitch class proposals rather than general discussions.

This research utilized association rule mining, a data mining technique designed to uncover relationships, patterns, and dependencies within a dataset. Association rules consist of an antecedent (or left-hand side) and a consequent (or right-hand side), with the antecedent identifying significant relationships between items and the consequent representing another item likely to co-occur. This approach has wide-ranging applications, including market basket analysis, customer behavior analysis, and recommendation systems.

In this study, tags were treated as items and dependencies as the co-occurrence of tags. The output of association rules provided insights into the frequency and strength of dependencies, measured through support, confidence, and lift. Support gauges how frequently a specific item set occurs, confidence quantifies the conditional probability of the consequent given the antecedent, and lift compares the likelihood of the consequent occurring given the antecedent.

This analysis proceeded in three steps: firstly, capturing the discussion\_ids of seed tags; secondly, creating a subset of comments with the same discussion\_id as the seed tags; and finally, computing association rules with support and confidence thresholds set at 0.001 to ensure the inclusion of both strong and weak associations. These rules were calculated on an aggregated weekly basis, offering snapshots of association patterns within the dataset over time.

The original dataset comprises 144,593 observations in total. During the analysis process, this study utilized data sourced from the discussion threads of proposal tags, which this study refer to as “seed tags.” Each discussion thread associated with glitch proposals contained an average of 4.41 (with a standard deviation of 6.31 and a median of 3) seed tags. As the number of tags in these discussion boards reflects the final curation, this study delved into the discussions centered around image subjects related to these seed tags to gain deeper insights into the evolution of tagging dynamics. For every proposal, this study first extracted the seed tags, then identified the discussion IDs of comments containing each seed tag. These discussion threads serve as platforms where volunteers engage in conversations regarding the presence of noise signals, often including additional tags beyond those initially submitted with the proposals. This broad approach was deliberately chosen due to uncertainties surrounding the current relationships among tags in the project.

Implementing this search strategy yielded a dataset comprising 61,657 comments (representing 42% of all comments) containing a total of 78,803 tags, out of which 4,219 were unique. Less than 12% of volunteers contribute comments in Gravity Spy, with 53% ( $N = 1,749$ ) of those volunteers included in our dataset. On average, each proposal contained 360 tags (with a standard deviation of 481 and a median of 80). The disparity between the average number of seed tags and those identified through our search indicates a challenge for volunteers in fully expressing the complete range of tags and their relationships. Each proposal involved an average of 167 volunteers (with a standard deviation of 313 and a median of 39).

To evaluate the thoroughness of proposers in identifying relevant tags, this study calculated a “hit rate,” which reflects the proportion of seed tags compared to the entire tag set from our search. The average hit rate stands at 13% (with a median of 5%). Although the hit rate may seem low, this study suspect that this proportion might be appropriate considering factors such as tags categorized as personomies (unique to a volunteer), misspellings, tags unrelated to noise characteristics in the image descriptions, or other phenomena.

Following association rule mining, the dataset expands to encompass 311,114 observations, which include 42 proposal IDs and a total of 610 pairs representing both left-hand and right-hand sides. Subsequently, a convergence analysis is conducted on the data derived from association rule mining. In this analysis, the focus lies on evaluating the support metric, which signifies the frequency or occurrence of specific itemset within the dataset.

## IV. METHODOLOGY

### A. MANN-KENDALL TEST

The Mann-Kendall test, a non-parametric statistical method, is employed to identify monotonic trends within time series datasets. Unlike parametric tests, it does not rely on assumptions of normality or linearity, rendering it robust across diverse data types. Monotonic trends signify consistent increases or decreases in a variable over time, without necessitating linear patterns. In the context of this test, the null hypothesis suggests an absence of monotonic trends within the data, while the alternative hypotheses encompass scenarios of upward, downward, or any form of monotonic trend.

In this study, the hypotheses are formulated as follows:

$H_0$ : No monotonic trend

$H_a$ : Downward monotonic trend

First, Arrange the data sequentially based on the order of collection over time, denoting the measurements as  $x_1, x_2, \dots, x_n$ , where  $x_i$  represents the measurement obtained at time  $i$  for  $i = 1, 2, \dots, n$ . Then, calculate the sign of all

possible differences  $x_j - x_k$ , where  $j > k$ . The number  $S$  is calculated through the formula in (1) where  $sgn(x)$  is defined in (2):

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^n sgn(x_j - x_k) \quad (1)$$

$$sgn(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \quad (2)$$

If  $S$  is a positive number, it indicates that observations obtained later in time tend to be larger than observations made earlier. Conversely, if  $S$  is a negative number, it suggests that observations made later in time tend to be smaller than observations made earlier. The variance of could be calculated through formula in (3) where  $n$  is the number of data points in the time series,  $m$  is the number of tied groups in the data,  $t_p$  is the number of data points in the  $p$ th tied group, and  $q_p$  is the number of data points minus the number of tied groups:

$$\text{Var}(S) = \frac{n(n-1)(2n+5) - \sum_{p=1}^m q_p(t_p-1)(t_p-2)}{18} \quad (3)$$

Finally, the test statistic  $Z_{MK}$  is calculated through the formula in (4):

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{\text{Var}(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sqrt{\text{Var}(S)}} & \text{if } S < 0 \end{cases} \quad (4)$$

The  $H_0$  is rejected if, at the significance level  $\alpha$ ,  $Z_{MK} \leq -Z_{1-\alpha}$ , and the trend could be considered as Downward monotonic trend. In this study, the significance level  $\alpha$  is established at 0.05.

### B. TEST FOR CONVERGENCE

While convergence is often equated with stationarity in various studies, its strict definition entails a time series gradually nearing a specific value with a continuous reduction in variation over time.

Some research endeavors to pinpoint this convergence by establishing a threshold, identifying the time point where the series fluctuates within this threshold. However, this approach encounters challenges as the optimal threshold can vary across different data types, posing a bottleneck to the methodology. Additionally, there is a risk of falsely claiming convergence, as the variance may not consistently diminish over time.

Thus, rather than pinpointing the convergence itself, this paper proposes an algorithm to detect the initiation of convergence, defined as the point where the time series' variance begins to decrease consistently. This approach aims to mitigate the uncertainty inherent in concluding



convergence by focusing on the observable onset of reduced variation in the time series.

---

**Algorithm 1:** FindConvergeStartPoint

---

```

Data: A time series  $L$ 
Result: The time point  $t$  where the convergence begins in  $L$ 
Input :  $L, \alpha$ 
Output:  $t$ 
 $t \leftarrow \infty$ ;
 $N \leftarrow \text{length}(L)$ ;
// Step 1: Calculate the Standard Deviations
for  $i \leftarrow 1$  to  $N - 1$  do
     $S[i] \leftarrow \text{var}(L[i : N])$ ;
end
// Step 2: Mann-Kendall Test,  $H_a$ : Downward monotonic trend
for  $i \leftarrow 1$  to  $N - 3$  do
     $p \leftarrow \text{Mann-Kendall}(S[i : N - 1])$ ;
     $\tau \leftarrow \text{Mann-Kendall}(S[i : N - 1])$ ;
    if  $p < \alpha$  and  $\tau < 0$  then
         $t \leftarrow i$ ;
        return  $i$ ;
    end
end
return  $t$ ;

```

---

The *FindConvergeStartPoint* algorithm is designed to provide a precise estimation of when convergence begins within a given time series, a crucial aspect in various analytical contexts. Its functionality relies on two primary inputs: the time series data itself and a user-defined significance level. Operating on these inputs, the algorithm systematically processes the time series data to determine the point at which convergence initiates.

The algorithm unfolds in a series of steps, each geared towards identifying the onset of convergence with rigor and efficiency. Initially, it initializes a variable to a theoretically infinite value, symbolizing the absence of any observed convergence point at the outset. Subsequently, it calculates the standard deviations of consecutive subsets of the time series data, starting from each data point and extending to the end of the series. This step captures the variability inherent in different time series segments, laying the groundwork for subsequent analysis.

Following the standard deviation calculations, the algorithm conducts a Mann-Kendall test on these computed standard deviations. The Mann-Kendall test, discussed in the previous section, evaluates whether a statistically significant downward trend exists in the standard deviations. Such a trend, if present, signifies a consistent reduction in variability over time, a hallmark of convergence in many contexts.

If the Mann-Kendall test yields a p-value below the specified significance level, indicative of a significant downward trend in standard deviations, the algorithm returns the time at which this trend commences. This identified time point is the estimated start of convergence within the time series data.

The *FindConvergeStartPoint* algorithm offers a methodical approach to discerning the incipient phase of convergence within a time series. By leveraging statistical tests and user-defined significance levels, it facilitates a nuanced

understanding of the evolving dynamics present in the data, empowering researchers to make informed interpretations and decisions based on the identified convergence points.

The program complexity of the *FindConvergeStartPoint* algorithm can be assessed in terms of time and space requirements. In terms of time complexity, the algorithm entails two primary loops, each iterating over the length of the input time series, denoted as  $N$ . The first loop calculates standard deviations for subsets of the time series, resulting in a time complexity of  $O(N^2)$ , as it iterates over each data point and computes the standard deviation for subsequent subsets. The second loop conducts the Mann-Kendall test, also iterating over the length of the time series, with each iteration involving a calculation based on the Mann-Kendall test, typically performed in linear time. Therefore, the overall time complexity of the algorithm remains  $O(N^2)$ .

Concerning space complexity, the algorithm necessitates additional space to store the standard deviations computed for each subset of the time series, resulting in a linear space complexity of  $O(N)$ . Moreover, a few auxiliary variables, such as  $t$ , require constant space.

## V. RESULTS AND DISCUSSION

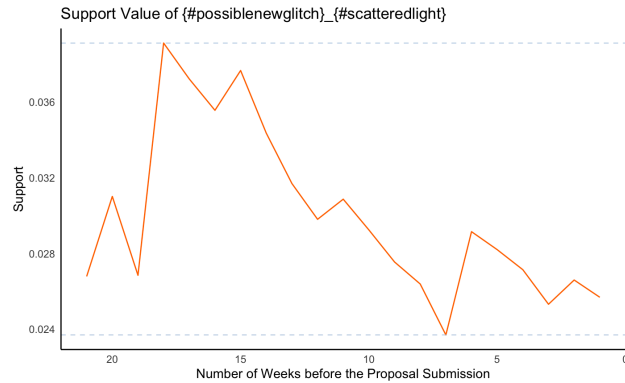
### A. START POINT OF CONVERGENCE

This study observed that 99.8% of support metric time series converge emphasizes the remarkable consistency of these metrics over time, signifying a robust tendency for tag pairs to stabilize in their relationships before the proposal submission deadline. This high convergence rate implies a predictable pattern in the behavior of tag pair relationships, instilling confidence in their reliability as indicators of collaborative dynamics.

Moreover, the average convergence start point occurring approximately 2.297225 weeks prior to the proposal submission reveals an intriguing trend of early stabilization in tag pair relationships. This early convergence suggests a proactive approach among collaborators, where adjustments and refinements to relationships are initiated well in advance of formal proposal submissions. This timeframe offers ample opportunity for stakeholders to fine-tune collaborative strategies, potentially optimizing research directions and enhancing the efficacy of joint efforts.

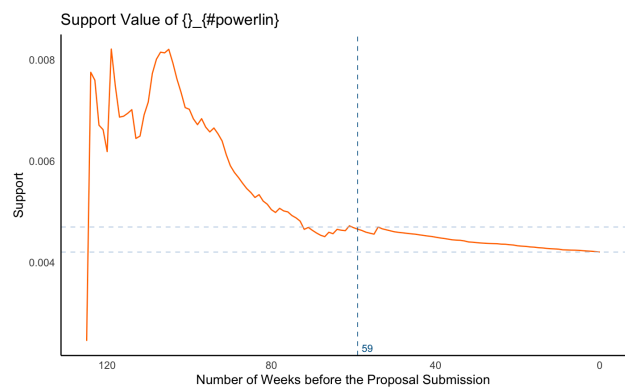
The early onset of convergence also implies a period of adaptation and refinement in tag pair relationships during the weeks leading up to proposal submission. This phase likely involves iterative discussions, resource allocation, and alignment of objectives, fostering a conducive environment for productive collaboration. By identifying this trend, researchers and stakeholders gain valuable insights into the dynamics of collaborative relationships, enabling them to anticipate and navigate the process of proposal submission with greater clarity and efficiency.

Overall, the high convergence rate and early stabilization of tag pair relationships highlight the reliability and predictability of collaborative dynamics, offering valuable guidance for effective research collaboration and proposal development.



**FIGURE 2.** The trend in association rule support for  $\{ \#possiblenewglitch \} \rightarrow \{ \#scatteredlight \}$  pair showing no trend of convergence.

To provide a visual representation of the proposed algorithm, **Figure 2** presents a support metric time series that has remained unconverged. Upon examination, no discernible pattern in variance or trend is evident throughout the entire time frame. This lack of convergence is visually depicted, emphasizing the absence of stabilization or consistency in the relationship dynamics captured by the support metric.



**FIGURE 3.** The trend in association rule support for  $\{ \} \rightarrow \{ \#powerlin \}$  pair showing a trend of convergence at 59 weeks before proposal submission.

In contrast, **Figure 3** illustrates a support metric time series that has converged precisely at 59 weeks before the proposal submission. A clear pattern emerges in the form of decreased variation across the entire time series, accompanied by a noticeable stabilization of values. This visual depiction vividly portrays the convergence of the support metric, highlighting the consistent and stable nature of the relationship dynamics captured within the data.

## B. STATIONARITY ANALYSIS

In the context of time series analysis, a stationary time series is one where statistical properties such as mean, variance, and autocorrelation structure remain constant over time. The focus of this study lies in assessing the stationarity of the “support” metric derived from association rule analysis of tag pairs. To accomplish this, the Augmented Dickey–Fuller test, a widely-used tool for testing the stationarity of a time series, is employed.

The methodology for pinpointing the time point at which stationarity is achieved involves a meticulous stepwise analysis of the support metric’s time series. The analysis begins from the initial appearance of the tag pair and proceeds incrementally until the proposal submission deadline. Initially, the Augmented Dickey–Fuller test is applied to evaluate the stationarity of the time series spanning this entire duration [17].

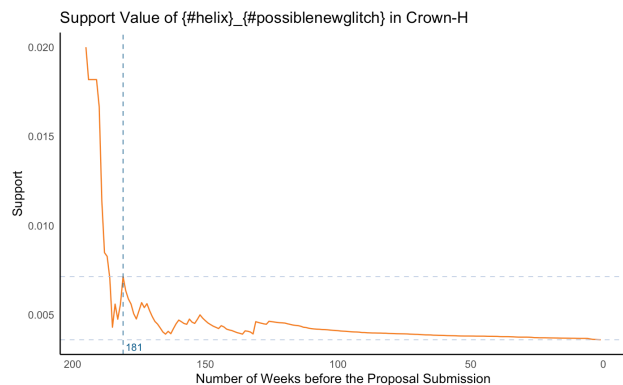
Should the initial analysis reveal non-stationarity, focus narrows to the period starting from the second week after the tag pair’s introduction and extends up to the proposal submission deadline. This iterative process continues, with each subsequent analysis commencing one week later than the previous one, until the penultimate week prior to the proposal submission deadline.

Throughout this iterative procedure, if the Augmented Dickey–Fuller test indicates a stationary time series for the support metric, the corresponding week number is meticulously recorded. Conversely, if the test consistently fails to identify stationarity across all iterations, it is deduced that the support for the tag pair does not attain stationarity within the observed timeframe. This systematic approach ensures a detailed exploration of the dynamics of the support metric and provides insights into its behavior over time.

The analysis undertaken unveils that 74% of tag pairs display stationarity before the proposal submission deadline. Among these stationary pairs, detailed examination reveals that the average timeframe for becoming stationarity stands at 100 weeks prior to the proposal submission, with the median convergence taking place at 96 weeks. These findings imply that a substantial majority of tag pairs stabilize in their relationships well in advance of the formal proposal submission process.

To effectively demonstrate the concept of convergence, a case study focusing on  $\{ \#helix \}_ \{ \#possiblenewglitch \}$  is conducted. Employing the methodology outlined earlier, the convergence point is observed at the 181st week preceding the proposal submission. The **Figure 4** depicts the comprehensive trend of the support metric spanning from 200 days prior to the proposal submission to the actual day of proposal submission.

In the **Figure 4**, the support value post-convergence exhibits fluctuations within the range of 0.003585347 to 0.007125891, with a variance of  $2.667543\text{e-}07$ . It becomes evident that the support value remains consistently stable around a specific value at the 181st week before the proposal submission. This detailed examination provides a clear illustration of how stationarity manifests within the context of tag pair relationships.



**FIGURE 4.** The trend in association rule support for {#helix} -> {#possiblenewglitch} pair showing a trend of stationarity at 181 weeks before proposal submission.

A discernible contrast between **Figure 3** and **Figure 4** reveals distinct characteristics of stationarity and convergence. Stationarity encompasses a phase where values may exhibit fluctuations, whereas convergence entails a gradual stabilization towards a specific value. While **Figure 3** depicts the transition towards convergence with a diminishing variability and an increasingly stable trend, **Figure 4** illustrates a lack of convergence, characterized by ongoing fluctuations without a discernible pattern of stabilization. This comparison underscores the significance of identifying convergence points, as they signify a pivotal moment where the dynamics of the time series become more predictable and stable.

## VI. CONCLUSION

In conclusion, this research offers valuable insights into the dynamics of tag pair relationships in the context of proposal submissions. Through meticulous analysis of support metric time series, it becomes evident that a significant majority of tag pairs exhibit convergence well in advance of the proposal submission deadline, indicating a propensity for stability in collaborative dynamics. The observation of high convergence rates and early stabilization points underscores the predictability and reliability of tag pair relationships, offering researchers and stakeholders valuable guidance in collaborative endeavors. Moreover, the visual representations provided by the support metric time series offer clear illustrations of convergence patterns, highlighting the importance of identifying and understanding these critical junctures in collaborative dynamics. By shedding light on the processes leading up to

proposal submissions, this research contributes to a deeper understanding of collaborative dynamics in research environments, ultimately facilitating more effective collaboration and proposal development strategies.

While the algorithm presented offers a structured method for pinpointing the convergence start point within a time series, it is imperative to acknowledge its inherent limitations. Primarily, the algorithm heavily leans on the Mann-Kendall test to discern downward monotonic trends in the time series. While widely utilized for trend detection, this test may not be universally applicable, potentially yielding inaccurate outcomes under certain circumstances such as non-linear trends or outlier presence. Additionally, the algorithm's efficacy is contingent upon parameter selection, including factors like the length of the sliding window for standard deviation calculations and the step size for traversing the time series. Poor parameter choices could compromise the precision and reliability of the identified convergence start point. Furthermore, the Mann-Kendall test's conservative nature may overlook conspicuous trends, further complicating accurate trend detection. Despite providing a structured framework for convergence analysis, the algorithm may lack robustness and scalability when confronted with extensive or intricate datasets. The algorithm's iterative nature and computational demands could render it impractical for large-scale or high-dimensional data analysis, posing significant challenges for real-world application.

While this study offers valuable insights, it is important to acknowledge its limitations. Firstly, the dataset may suffer from sampling bias, potentially limiting the generalizability of the findings to a broader population or context. Additionally, there may be instances of incomplete or missing data, which could introduce gaps in the analysis. The quality and accuracy of the extracted tags from discussions may vary, influenced by inconsistencies, misspellings, or ambiguities introduced by volunteers. Moreover, volunteer behavior within the discussions could introduce biases, with certain individuals being more active or influential than others. It's also essential to recognize that correlation does not imply causation, and further investigation would be needed to establish causal relationships. The study may not fully capture the evolution of tagging practices over time, and the analysis may be constrained by the scope and depth of the research. Furthermore, the findings may be context-specific and may not be applicable to other domains or settings. By acknowledging these limitations, the study can provide a more nuanced understanding of its findings and guide future research efforts to address these gaps.

## ACKNOWLEDGMENT

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. J. Ling wishes to express profound gratitude to Dr. Corey B. Jackson, the esteemed honor thesis advisor, whose unwavering support, insightful guidance, and profound expertise have been invaluable throughout this journey. J. Ling is deeply appreciative of Carmela Diosan, the Academic Advising Manager, whose assistance with administrative tasks has been indispensable. Heartfelt thanks are extended to J. Ling's beloved family for their boundless love, encouragement, and sacrifices, which have been the cornerstone of strength and motivation. Equally, gratitude is extended to friends and peers whose unwavering encouragement and support have been a constant source of inspiration. Furthermore, heartfelt acknowledgment is given to the University of Wisconsin-Madison, Department of Statistics, and the University of Wisconsin-Madison, The Information School, for their unwavering commitment to providing the necessary resources and conducive environment for this research endeavor. J. Ling is also deeply grateful to the Purdue University Summer Undergraduate Research Fellowship (SURF) program for providing invaluable opportunities for research exploration and growth. Special appreciation is extended to Dr. Konstantina Gkritza and Dr. Xiaodong Qian, J. Ling's first research mentors. Their mentorship, guidance, and unwavering belief in their potential ignited the spark that set J. Ling's research journey in motion. Their dedication to fostering an environment of intellectual curiosity and scholarly development laid the groundwork for a lifelong pursuit of knowledge and inquiry. Each of these contributions, whether through guidance, support, or resources, has played an integral role in the completion of this thesis. For these countless gestures of kindness, encouragement, and support, J. Ling is genuinely thankful. This thesis stands as a testament to the collective effort and collaboration of all involved, and their contributions will forever be remembered with deep appreciation and gratitude.



## REFERENCES

- [1] S. Faraj, G. Von Krogh, E. Monteiro, and K. R. Lakhani, "Special Section Introduction—Online Community as Space for Knowledge Flows," *Inf. Syst. Res.*, vol. 27, no. 4, pp. 668–684, Dec. 2016, doi: 10.1287/isre.2016.0682.
- [2] S. Faraj, S. L. Jarvenpaa, and A. Majchrzak, "Knowledge Collaboration in Online Communities," *Organ. Sci.*, vol. 22, no. 5, pp. 1224–1239, Oct. 2011, doi: 10.1287/orsc.1100.0614.
- [3] McGill University, S. Faraj, S. Kudaravalli, HEC Paris, M. Wasko, and University of Alabama at Birmingham, "Leading Collaboration in Online Communities," *MIS Q.*, vol. 39, no. 2, pp. 393–412, Feb. 2015, doi: 10.25300/MISQ/2015/39.2.06.
- [4] F. Ricci, K. Bontcheva, O. Conlan, and S. Lawless, Eds., *User Modeling, Adaptation and Personalization: 23rd International Conference, UMAP 2015, Dublin, Ireland, June 29 -- July 3, 2015. Proceedings*, vol. 9146. in Lecture Notes in Computer Science, vol. 9146. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-20267-9.
- [5] M. Osinski and N. Rummel, "Towards Successful Knowledge Integration in Online Collaboration: An Experiment on the Role of Meta-Knowledge," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–17, Nov. 2019, doi: 10.1145/3359133.
- [6] D. Diriker, M. Boons, P. Tuertscher, and A. J. Porter, "Taking on Grand Challenges Through Collaborative Crowdsourcing: The Importance of Common Ground for Knowledge Integration," *IEEE Trans. Eng. Manag.*, pp. 1–14, 2024, doi: 10.1109/TEM.2022.3191292.
- [7] I. Peters and K. Weller, "Tag Gardening for Folksonomy Enrichment and Maintenance".
- [8] J. Makani and L. Spiteri, "The Dynamics of Collaborative Tagging: An Analysis of Tag Vocabulary Application in Knowledge Representation, Discovery and Retrieval," *J. Inf. Knowl. Manag.*, vol. 09, no. 02, pp. 93–103, Jun. 2010, doi: 10.1142/S0219649210002553.
- [9] S. Kopeinik, E. Lex, P. Seitlinger, D. Albert, and T. Ley, "Supporting collaborative learning with tag recommendations: a real-world study in an inquiry-based classroom project," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, Vancouver British Columbia Canada: ACM, Mar. 2017, pp. 409–418. doi: 10.1145/3027385.3027421.
- [10] E. Santos-Neto, M. Ripeanu, and A. Iamnitchi, "Tracking User Attention in Collaborative Tagging Communities".
- [11] U. M. Fayyad, *Advances in knowledge discovery and data mining*. Menlo Park (Calif.): AAAI press MIT press, 1996.
- [12] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme, "Mining Association Rules in Folksonomies," in *Data Science and Classification*, V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, Eds., in Studies in Classification, Data Analysis, and Knowledge Organization. , Springer Berlin Heidelberg, 2006, pp. 261–270. doi: 10.1007/3-540-34416-0\_28.
- [13] J. A. Diaz-Garcia, M. D. Ruiz, and M. J. Martin-Bautista, "A survey on the use of association rules mining techniques in textual social media," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1175–1200, Feb. 2023, doi: 10.1007/s10462-022-10196-3.
- [14] M. Simionescu, W. Strielkowski, N. Schneider, and L. Smutka, "Convergence behaviours of energy series and GDP nexus hypothesis: A non-parametric Bayesian application," *PLOS ONE*, vol. 17, no. 8, p. e0271345, Aug. 2022, doi: 10.1371/journal.pone.0271345.
- [15] G. A. Carlino and L. O. Mills, "Are U.S. regional incomes converging?," *J. Monet. Econ.*, vol. 32, no. 2, pp. 335–346, Nov. 1993, doi: 10.1016/0304-3932(93)90009-5.
- [16] Y. Cao *et al.*, "Visualizing Collective Idea Generation and Innovation Processes in Social Networks," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 5, pp. 2234–2243, Oct. 2023, doi: 10.1109/TCSS.2022.3184628.
- [17] D. A. Dickey and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series With a Unit Root," *J. Am. Stat. Assoc.*, vol. 74, no. 366, p. 427, Jun. 1979, doi: 10.2307/2286348.