Web as an Evolutionary Ecosystem: Emergence of Keystone Species

Palin Choviwatana1, Shota Ejima3, Mizuki Oka1,3 and Takashi Ikegami2,3

1University of Tsukuba, Tsukuba, Japan
2The University of Tokyo, Tokyo, Japan
3Alternative Machine Inc. Tokyo, Japan
mizuki@cs.tsukuba.ac.jp

Abstract
Research is being conducted to understand human social interactions on the Web as a biological ecosystem. Keystone species in a biological ecosystem are defined as species that significantly impact the ecosystem if removed, irrespective of its low biomass. Identifying keystone species is an important issue, as they play a vital role in maintaining the entire ecosystem and its biodiversity. We hypothesize that a Web system is akin to an open, living ecological system that evolves and sustains itself by constantly updating its elements, which are sustained by the emergence of keystone species. We use data from an online bulletin board and identify keystone threads (“species”) that have a large impact if they are removed or become unpopular, despite their small population size. Our analysis confirms that keystone threads do exist in the system. The system seems to asymptotically evolve to a critical state. At the same time, the number of keystone species increases, and metabolism is enhanced. From a network topological perspective, the system evolves into a network with a high degree, closeness, and PageRank centralities. These findings suggest that keystone species play an important role in the evolution of online ecosystems. Further, by having keystone species, the system itself can decrease stability and bring about diversity to the ecosystem; consequently, the system can evolve.

Introduction
People’s communication in online communities creates enormous log data on a daily basis, and their interactions create new ideas that spread through social networks. Richard Dawkins stated that influence and information spread as memes analogous to genes (Dawkins, 1976). This meme evolution theory has recently regained attention, as large numbers of traces of individual actions and remarks are constantly recorded and accumulated on the Web, after which they can be quantitatively analyzed (Leskovec et al., 2009; Segev et al., 2015). Furthermore, such an evolutionary biological perspective and its mathematical framework have been proposed for online community analysis, such as social media (Beutel et al., 2012; Matsubara et al., 2015; Adamic et al., 2016; Li et al., 2019). As social media matures, the number of users grows, and the structure of the service itself often changes, with its contents becoming increasingly varied (Secretan et al., 2011; Ikegami et al., 2017) as if the system is biologically evolving. These studies suggest that an analogy can be borrowed from biological systems, and if correctly applied to social systems, it can open new perspectives for analyzing and understanding social systems.

In biological ecosystems, the concept of dominant and keystone species exists to define influential species (Power et al., 1996). Dominant species are defined as species having both high biomass and a large impact on the ecosystem. In contrast, keystone species have a large impact on the ecosystem despite their low biomass in proportion to the ecosystem. Figure 1 depicts the relationship between dominant and keystone species with the abundance of species and their impacts on ecosystems. Originally, the concept of keystone species was coined after ecologists removed a starfish species from its ecosystem. It was observed that when the keystone species was removed, despite its small population size, the balance of the ecosystem collapsed, leading to the extinction of many other species (Paine, 1966). Many experiments on biological ecosystems have demonstrated that...
less abundant species also have strong effects on communities and ecosystems (Paine, 1969; Estes et al., 2016), and finding and protecting those species is vital to maintain the entire ecosystem and its biodiversity. In contrast to dominant species, keystone species are considered difficult to identify, as their population is small. Moreover, although keystone species are known to exist in biological ecosystems, their characteristics are not fully understood (Jordán et al., 2006; Jordán, 2009; Genrich et al., 2017; González et al., 2010; Mello et al., 2015). Identifying and analyzing keystone species in a Web system allow us to obtain insights into their key characteristics in biological systems.

In online social communication systems, identifying keystone species in ecosystems can be regarded as recognizing influential users or content, which have been extensively studied in the literature. These studies include identifying influential users based on follow-follower networks ( Kwak et al., 2010) or diffusion models of the spread of influence and information through social networks (Ver Steeg and Galstyan, 2012; Iwata et al., 2013). The influential users or contents obtained from these studies can be regarded as identified dominant species in the biological ecosystem, but to the best of our knowledge, only a few studies have discussed keystone species in the context of online communities. Therefore, in this study, we introduce the concept of keystone species in an online communication system and hypothesize that a web system is akin to an open, living ecological system that evolves and sustains itself by constantly updating its elements, which is sustained by the emergence of keystone species.

Specifically, we apply the concept of keystone species to online bulletin boards by considering a bulletin board as a social ecosystem, and a thread as a species. Here, the removal of threads can be regarded as a state wherein there is no attention or action by users on the thread. On an online bulletin board, users create threads related to specific topics and themes to communicate. Threads can be said to influence each other because users move back and forth among threads. In other words, threads are in a competitive relationship to obtain a limited number of users’ resources (e.g., time). Such an environment can be considered similar to the competitive relationship between species in biological ecosystems.

In contrast to the standard biological definition of keystone species, we extend our definition to a more dynamic and network-based one (Ikegami, 2005). This means that we treat the keystone species as time-varying and context-dependent (i.e., dependent on the combinations of other species) processes. We model each thread as an event sequence of all the actions performed by users (i.e., views, comments, replies, and claps) and estimate the influence among threads by using the multivariate Hawkes process (Hawkes, 1971), which is a self-exciting stochastic point process. The Hawkes process was originally used to analyze earthquakes (Ogata, 1988) before being applied to many other fields, such as financial market modeling (Filimonov and Sornette, 2012). In recent years, it has been used for influential user identification (Zhou et al., 2013) and recommendations (Zarezade et al., 2017), user interest modeling (Lan et al., 2018), and steering user activity by controlling dynamics (Farajtabar et al., 2016) on social media.

To identify keystone species, it is necessary to measure the degree of impact when the species is removed from the system. Gupta et al. (2018) proposed a method for characterizing the effect of removing selected events in the history of the Hawkes process. We employ a similar method in this study to simulate the degree of impact caused by thread deletion. Specifically, we use the fitted parameters of the multivariate Hawkes process and simulate the model to compute the degree of impact caused by deleting each thread from the system. This is done by computing the activity level of the entire bulletin board (Onaga and Shinomoto, 2016), comparing the activity levels from before and after the removal, and regarding the difference as the impact of the removal of each thread. If a thread has a high degree of impact despite having a small density (i.e., the number of users participating in the thread), then we identify the thread as a keystone thread.

The contributions of this study are summarized below.

- We introduce the biological concept of keystone species to a Web system (i.e., online bulletin board data) and demonstrate the existence of keystone species on the Web.
- We analyze the found keystone species and their interactions in the network to understand their characteristics.
- We show that a Web system is akin to a living ecosystem that evolves and continually updates itself by replacing the old with the new, a process which is sustained by the emergence of keystone species.

**Definition of Keystone Species**

**Keystone Species in Biological Ecosystems**

In a biological ecosystem, a keystone species is defined as a species whose impact on its ecosystem is substantial and disproportionately large relative to its abundance (Power et al., 1996). A classic example of a keystone species is a certain species of starfish in California. When a low-biomass starfish (the predator in its ecosystem) was removed, the balance of the ecosystem collapsed, causing the extinction of other species in the ecosystem. Thus, the keystone species is considered to be an essential species in maintaining the stability of the entire ecosystem. In Power et al. (1996), the presence of keystone species was identified by a predator species removal, historical data on species removals, predator re-introductions, and comparative studies on relatively high-profile species.

To define keystone species, the population density $p_i$ occupied by species $i$ in the ecosystem and the impact $I_i$ when...
Keystone Species in a Web System

We apply the concept of keystone species to a Web system (specifically online bulletin boards), considering a bulletin board as a social ecosystem, and a thread as a species. The population size of a thread is computed using the number of user actions such as comments and claps in the thread; these activities keep the species alive. Given the threads and event sequences of actions, the challenge is to estimate the impact of a thread when it is removed. On the online bulletin board community, threads can be said to influence each other because users move back and forth among threads as discussed above. We use the multivariate Hawkes process to estimate the degree of influence between threads, create an influence thread network, and simulate the thread removal. The Hawkes process (Hawkes, 1971) is a self-exciting stochastic point process and a probabilistic model that considers the influence of past event occurrences on future event occurrences. The multivariate Hawkes process is a mutually exciting version of the Hawkes process. By using this procedure, it is possible to calculate the magnitude of influence between two or more sequences. Then, we calculate the impact on the online bulletin board when each thread is deleted using the estimated degree of influence. Using this impact value and the density defined for each thread, a thread that has a high overall impact when deleted, despite its small density, is identified as a keystone.

The intensity function \( \lambda_i(t) \) of the event sequences at time \( t \) of thread \( i \) is represented as follows:

\[
\lambda_i(t) = \mu_i + \sum_{j=1}^{D} \sum_{t'_j < t} \phi(t - t'_j)
\]

where \( D \) is the number of threads, \( \mu_i \) is the base intensity of thread \( i \), and \( t'_j \) represents the \( k \)-th event sequence of thread \( j \). \( \phi(t) \) is called the kernel function, and we use the exponential kernel function \( \phi(t) = \alpha_{ij} e^{-\beta_{ij} t} \). \( \alpha_{ij} \) represents the degree of the influence of thread \( j \) on thread \( i \), and \( \beta_{ij} \) represents the decay rate of the intensity of thread \( i \) raised by the event sequences in thread \( j \). These parameters can be analytically obtained by the maximum likelihood method, given an event sequence (Ogata, 1988).

We define threads as nodes, set the estimated parameter \( \alpha_{ij} \) as the edge representing the degree of influence from thread \( i \) to \( j \), and construct a network. We consider this thread network as an ecosystem and identify a keystone thread that plays an important role in maintaining the stability of the network, i.e., stable communication within the online bulletin board.

Impact of Removal of a Thread

To calculate keystone threads, we need to calculate the impact \( I_i \) on the ecosystem, viz., the thread network, when thread \( i \) is removed. Here, we use a cascading condition \( C \), which represents the activity level of the network (Onaga and Shinomoto, 2016). Let \( \alpha = \{\alpha_{ij}\}_{i,j=1,\ldots,D} \), \( \mu = \{\mu_i\}_{i=1,\ldots,D} \) be the parameters estimated via the multivariate Hawkes process. Then \( C \) is defined as follows:

\[
C = \frac{\sum_{i,j}(\Lambda \Lambda L^T)_{ij}}{\sum_i(\lambda_i)},
\]

where \( L = \sum_{n=0}^{\infty} \alpha = (I - \alpha)^{-1}, \Lambda = \text{diag}(\lambda) = \text{diag}(\{\lambda_i\}_{i=1,\ldots,D}) = \text{diag}(L \mu) \). The value of \( C \) becomes higher as the event sequence of each node (thread) of the network becomes more bursty. We define this burstiness to represent the activity level of the network. Onaga and Shinomoto (2016) also theoretically showed that when the value of the cascading condition \( C \) is greater than two, the event occurrence of the entire network undergoes a phase transition to a bursty, non-stationary state. Conversely, if the
value of $C$ is less than two, the event occurrence state of the network is a stationary state.

The change in activity level of the network $\Delta C_i$ when deleting a certain thread $i$ in the network is defined as follows:

$$\Delta C_i = \sum_{j \neq i} (C - C'_j),$$

where $C'_j$ is the cascading condition calculated by setting the influence between threads $i$ and $j(\neq i)$ (i.e., $\alpha_{ij}, \alpha_{ji}$) to 0. Performing the same calculation for all $j \neq i$, and calculating the differences with the original $C$ and all the $C'_j$, we obtain $\Delta C_i$ by adding the differences. This $\Delta C_i$ is equivalent to $t_N - t_D$ in the equation (1). We can thus define the impact $I_i$ by removing the thread $i$ from the network as follows:

$$I_i = \frac{t_N - t_D}{t_N} = \frac{\Delta C_i}{C}$$

**Experiments**

The data we used were provided by QON Inc., which is an online community service in Japan. On this online bulletin board service, users can communicate with one another by posting threads and replying to the threads by commenting and clicking claps (corresponding to likes) for threads and/or comments. More than 100 companies use this service as an online bulletin board for communication between themselves and their consumers, and the total number of users exceeds one million. Each company is provided with a community wherein the company can run its own bulletin board. We analyzed 15 communities among them from January 2014 to May 2019. Note that these 15 communities are different from each other in terms of community purposes, community life periods, user engagement rates, and groups of members. Thus, we considered each community as a different ecosystem and created a thread network separately for each community using corresponding community data.

We analyzed the event sequences composed of user actions, such as views, comments, and claps, on threads. The event sequences were divided into two-week windows, where 839 windows were obtained for the period between January 2014 and May 2019, and the parameters were fitted using the multivariate Hawkes process for each window.

We disregard threads having less than 50 actions in the event sequences, as fitting cannot be performed if the number of actions is too small. We also disregard communities having less than 15 threads, as analysis of the network is not meaningful if the number of threads (nodes) is too small. Accordingly, 10,147 unique threads, 839 windows, and 15 communities are targeted for our analysis. Table 1 shows the statistics of the data of the 15 communities of the online bulletin board.

The temporal development of the number of actions over time in each community is depicted in Figure 2. In the figure, one row represents one community, and a sequence of dots in one row represents an event sequence of actions over the two-week window. The color indicates the number of threads with light yellow being low and dark blue being high. Various patterns of actions are observed, depending on communities; some communities increase their numbers of threads over time, while others decrease theirs and eventually fade away from the platform.

![Figure 2: Window distribution over time for all 15 communities, with a dot representing one two-week window. The colors indicate the number of threads, ranging from light yellow (low) to dark blue (high).](http://direct.mit.edu/isal/proceedings-pdf/isal2020/32/230/1908574/isal_a_00263.pdf)

For the population density, which we need to identify to calculate keystone threads, we use the number of user actions in the thread. Let $N$ be the number of target threads and $u_i$ be the number of total actions in thread $i$ in a window. Then the population density $p_i$ of thread $i$ can be calculated as follows:

$$p_i = \frac{u_i}{\sum_{j=1}^{N} u_j}.$$  

(8)

The threshold coefficients $c_p$ and $c_I$ are set to 0.25 and 0.5, respectively, with reference to Ikegami (2005). This means that the thread $i$ whose $p_i$ is less than one-fourth of the maximum population thread $c_{max}$ and $I_i$ is more than half of the maximum impact thread $I_{max}$ in a window is identified as a keystone thread. It is noteworthy that thresholds are computed in relation to the maximum values of population and impacts, respectively, and the threshold values are different for each window.
Results

Existence of Keystone Species in a Web System

First, we show the temporal development of the identified keystone species of 15 communities in Figure 3-(a). Each dot represents a two-week window and is colored by the proportion of keystone threads in the window from light yellow to dark blue, with the latter representing a higher proportion of keystone threads in the window. An example of a keystone identification result in a window (marked in the black square in Figure 3-(a)) is depicted in Figure 3-(b) as the scatter plot with $p_i$ on the horizontal axis and $I_i$ on the vertical axis. The dashed line represents $y = x$, indicating a proportional relationship between population and impact. Keystone threads are plotted in red circles. The corresponding thread network of the window is depicted in Figure 3-(c).

Among the 15 communities, 1,666 unique threads were detected as keystones in 222 windows (26% of target windows) in total. Of the 1,666 unique keystone threads, 85% of them (1,419 threads) were detected as keystone threads only once, 12% of them (198 threads) were detected twice, and the maximum number of keystones detected was nine. In total, 2,015 nodes were detected as keystone threads. There were two communities in which keystone threads were not identified (communities 10 and 15). These results confirm the existence of keystone species in a Web system. We then analyze the found keystone species and the thread networks to understand their characteristics.

Analysis of the Keystone Species

In literature, network theory is applied to help keystone identification and characterization, since network metrics can represent a degree of node importance from a network topological perspective. Several studies have attempted to identify the traits of keystone species by modeling a biological ecosystem as a network, a species as a node, and a species relationship/interaction as an edge (Jordán et al., 2006; Jordán, 2009). Research results showed that there is a strong tendency for keystone nodes to have high scores in centrality metrics (Genrich et al., 2017; González et al., 2010; Mello et al., 2015). For example, González et al. (2010) examined keystone species in pollination networks and showed that keystone species have high degrees in closeness centrality and betweenness centrality. A high closeness value means that the node is at the center of the network being close to many other networks, and a high betweenness value means that the node acts as a connector among sub-networks. Another study (Mello et al., 2015) investigated keystone species across seed dispersal networks and found that keystone species in the networks were the only ones to attain the highest scores of closeness centrality and accessibility.

These studies collectively provide important insights into the characteristics of keystone species on different topological views and different ecological networks. Further, one may suppose that centrality metrics and other impor-
Figure 5: Temporal development of cascading condition value $C$ that indicates the activity level of the thread network. Each line depicts a community with a dot representing a window. The window containing a keystone thread is colored in red.

Figure 6: Distributions of windows with keystone and cascading condition values of all communities over their lifetimes. The temporal development of the distributions is depicted every six months from left to right. The X-axis shows the cascading condition value and the Y-axis displays the number of windows with keystone threads. The blue bar indicates a bin that contains the value 2. The dashed line represents a cumulative number of all the windows.

tant topological indices could characterize keystone species. Accordingly, we also apply network analysis to characterize keystone species in our thread networks. Specifically, we investigate the networks by using four common centrality metrics, namely, degree, closeness, PageRank, and betweenness centralities. Degree centrality indicates how likely the node is to be a hub in the network, and PageRank indicates how much information or influence tends to gather at the node when it spreads over edges.

Centrality values are computed for all nodes in each network and ranked in descending order for normalization to compare among different networks. Figure 4 shows the distribution of keystone nodes for each of the centrality metrics. Centrality values on each network are ranked on the x-axis and the number of keystone nodes is shown on the y-axis. The dashed line represents the cumulative number of all the nodes. We can observe that the keystone nodes tend to have a high ranking of degree and closeness centralities compared to the other non-keystone nodes. This implies that the keystone nodes tend to connect closely with many other nodes, being the hubs in the network. There is also a slight tendency for keystone nodes to have relatively high values in PageRank centrality. However, the betweenness centrality values of the keystone nodes are distributed rather equally in the ranking, with a peak in the lower ranking. This implies that a keystone node may be a connector between different parts of the network or otherwise, meaning there may be a shortcut that does not go through the hub between the subgraphs that the hub mediates. Because betweenness is sensitive to the shortest path, if there are other paths that are slightly shorter, then all the information goes through the shorter paths. Overall, keystone species in a web system tend to have a high degree, closeness, and PageRank centralities, and these results are consistent with existing biological research.

Analysis of Communities

The number of keystone threads differs among communities; some communities have larger numbers of keystones and others have no keystone threads (communities number 10 and 15). Thus, we are interested in understanding the characteristics of communities wherein keystone threads
tend to appear.

Figure 5 shows the temporal development of the cascading condition value $C$ that indicates the activity level of the thread network. Each line depicts a community with a dot, representing a thread network of a window. A window containing a keystone thread is colored in red. Interestingly, networks in which keystones appear have relatively low cascading condition values, with a median value of 2.2. In contrast, the median cascading condition value of networks in which a keystone does not appear is 7.8. Recall that the system is in a critical state when the cascading condition value is equal to 2. Looking at the temporal change in the $C$ value in the distribution for every six months, as shown in Figure 6, we can observe that there is a peak in the $C$ value around 2 (colored in blue) in all the time periods. The height of the peak gradually increases as the community evolves and decreases as it decays. These results suggest that the Web system evolves asymptotically to a situation whereby the condition value $C$ is equal to the critical value 2. This phenomenon of a Web system evolving towards a critical state or to the edge of chaos has been well observed in other Web systems, such as in the congestion rate of packet-switched networks (Oka et al., 2015) or the number of user actions of a social network service (Ikegami et al., 2017).

We are then interested in conducting analysis in terms of the “metabolism” of the community. We investigate how keystone appearance is related to the process of updating nodes in the community by using two metrics, namely, node birth rate and node death rate. The former is a proportion of newly created nodes (threads) in the community, while the latter denotes a proportion of threads that will no longer appear in the community. These metrics are computed on each thread network over a community and their averages are also obtained. Figure 7 depicts the relationship between node birth rate and death rate of 15 communities colored by the average ratio of keystone threads in networks. It shows a tendency that the more keystone threads are contained in the community, the higher the birth and death rates.

This tendency is confirmed by computing the Pearson correlation coefficient between birth (or death) rate and the keystone ratio in networks is shown in Figure 8. Each dot in the figure represents each community with birth (or death) rate on the x-axis, and the ratio of networks with keystone nodes is presented on the y-axis. There are moderately positive and significant correlations between the ratio of keystone nodes and the birth rate ($r = 0.57$, $p = 0.027$). The correlation with death rate is positive but not significant ($r = 0.54$, $p = 0.092$). This implies that the communities in which keystone threads frequently appear are constantly changing—being renewed—and the old is being replaced with the new.

**Conclusion**

In this study, we introduced the concept of keystone species (an important species for maintaining the stability of the entire ecosystem and its biodiversity) to the analysis of human online communication, using an online bulletin board. We considered threads in online bulletin boards as species in a biological ecosystem, and modeled influences among threads by using the multivariate Hawkes process. We constructed a thread network with threads as nodes, estimated influence as edges, and calculated the degree of impact on the activity level of the entire network when a thread is removed. Accordingly, we showed that there are threads that play keystone roles. We found several factors that characterize keystone species in a Web system. First, an analysis of threads from the viewpoint of network centrality metrics shows that keystone threads tend to have high degree, closeness, and PageRank centralities compared to other non-
keystone threads in the same network. Second, keystone species appear frequently when the system is in a critical state. Third, keystone threads tend to appear frequently in networks wherein the old threads are replaced with new ones with higher birth and death rates.

If a system is too stable, the evolution or emergence of a new species may not occur, and its diversity may be lost. Our findings suggest that keystone species play an important role in the evolution of ecosystems; by having keystone species, the system itself can decrease stability and bring about diversity to the ecosystem, and the system can evolve as a result.

Keystone species play an important role in preserving the stability of biological ecosystems. When considering an online community as an ecosystem, its stability is realized by its activity and engagement by users. Identifying content that may work to increase the activity of the community despite having a small population or low event occurrence is difficult in existing methods, so the proposed method and analysis give a new perspective for the stable operation of a Web system.

Acknowledgements
The authors would like to thank QON Inc. for providing the data for this study. This work was supported by JSPS KAKENHI Grant Number JP19H04214.

References


