Understanding the Structural Characteristics of Data Platforms Using Metadata and a Network Approach

TERUAKI HAYASHI AND YUKIO OHSAWA
Department of Engineering, The University of Tokyo, Tokyo 113-8656, Japan

Corresponding authors: Teruaki Hayashi (hayashi@sys.t.u-tokyo.ac.jp) and Yukio Ohsawa (ohsawa@sys.t.u-tokyo.ac.jp)

This work was supported by a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO), in part by the MEXT Quantum Leap Flagship Program (MEXT Q-LEAP) under Grant JPMXS0118067246, in part by Kyodo Printing Company, Ltd., and in part by the Artificial Intelligence Research Promotion Foundation.

ABSTRACT With the emergence of global platforms for trading and buying/selling data, data have become a profitable commodity. The growth of such platforms has necessitated the further expansion of the scope of data in digital economies. To this end, understanding the nature of available data and their relationships between them has become an important challenge for expanding their use. Thus, in this study, we assumed data on the platforms as a population and metadata as the samples. Thus, we quantitatively investigated the structural characteristics of data platforms, while avoiding the risk of lost business opportunities and privacy issues by not sharing the data themselves. By observing the characteristics of data and variables, we found that the data network had a structure that was locally dense and globally sparse, which is quite similar to networks of human relationships. Moreover, we found that data play different roles on the platforms when divided into sharing conditions, namely, shareable data and sensitive data. Finally, we discussed the potential tactics for individuals who create/use data platforms based on our findings. The contributions of this study include a new framework for data platform observation and a method that uses metadata and a network approach to analyze structural characteristics of data.

INDEX TERMS Data marketplace, data platform, data utilization, metadata, network analysis.

I. INTRODUCTION

In recent years, technology has evolved to allow a wide variety of data to be generated every moment, and new types of data are continually being introduced. Data exchange has led to the discovery of new intra- and cross-disciplinary knowledge [1], [2], and data have become a transferrable and exchangeable resource. Furthermore, various initiatives have been launched to solve complex problems by exchanging and combining data from different domains for purposes other than their originally intended uses [3], [4]. Different forms of data markets, including secondary data and data transaction markets [5]–[8], and related services, have emerged as platforms to conduct these transactions [9]–[11]. Data from various stakeholders are present in the data markets of “data platforms”—business players who develop online marketplaces where companies that want to buy and sell data (buyers and suppliers) from different areas can participate freely.

The associate editor coordinating the review of this manuscript and approving it for publication was Yong Xiang.
discussing rational strategies for assessing data are important challenges that must be overcome to expand the use of data platforms in the digital economy. To understand the aspects and interactions of data on platforms, it is necessary to develop an accurate understanding of the structure of data and their relationships on the platform. In other words, not only is the analysis of the individual pieces of data important, it is also important to investigate the structural characteristics of the population of data. Accordingly, data about data, i.e., metadata, are an effective analysis target. In this study, we assumed the data traded on platforms to be the population and the collected metadata to be samples. Thus, we quantitatively investigate the structural characteristics, relationships, and combinability of data on the platforms.

This paper is the first to discuss the use of metadata from different domains and the analysis of the variables and data characteristics on data platforms. This research is novel in that it enhances understanding of the structural characteristics of data populations on platforms using metadata and a variable-based network approach. To date, there have been studies on data platforms and exchanges based on the game theory [16], business roles and marketability [17], [18], market models [6], trading and pricing models [19], [20], tokenization of data using blockchain based systems [11], data protection and digital rights management models [8], [21], and privacy issues [22], [23], but none based on variables in the data and a network approach. Moreover, we focus not only on shareable data, which are typical of open data, but also sensitive data that generally cannot be shared, such as the private data of corporations and individuals. In prior studies, only networks of open data, such as Linked Open Data were investigated [24], [25]. In addition, ontology mapping has also been performed to investigate exchanges between and combinations of different data pairs [26], [27]. Notably, data platforms are not places where superficial information about data is displayed nor do they only manage shareable data. Instead, data platforms are dynamic systems that include the interactions of both shareable and sensitive data. However, the connections between data that are not generally shareable have not been addressed adequately thus far. Considering the complementarity and connections between data, it is likely that there is a community similar to a network of human relationships. Therefore, to find the characteristics of the population and the relationships between the data on the platforms is a novel approach. The contribution of this study is a quantitative method to investigate the structural characteristics of data using network analysis with the conditions for sharing data on data platforms.

The remainder of this paper is organized as follows. We explain our approach, method, and dataset in Section 2. In Section 3, we discuss the characteristics and distributions of the data and variables. In Section 4, we describe the creation of the data networks and discuss their characteristics with regards to data sharing conditions. Based on the characteristics of the data networks, we describe several tactics for data platformers in Section 5. Finally, the conclusions and future work are summarized in Section 6.

**II. APPROACH**

**A. RESEARCH QUESTIONS AND METHOD**

In this study, we interpreted the macro-level law of data, which is the common rule behind data when a population is observed. Even when not designed and acquired in a unified manner with common rules, or acquired separately for different purposes, data share similarities [28]. In this regard, although the micro-level features that characterize individual data are important, it is also essential to understand the macro-level law governing the detailed features of the data. To quantitatively examine the structural characteristics of data on platforms, data on those platforms were assumed to be the population and the collected metadata to be the samples. It is important to understand the structure of data across various fields with variables as the focus, because variables are frameworks for data. A variable is a logical set of attributes, and an attribute (or value) is a characteristic of a person or a thing [29]. For example, the variable “sex” has the attributes of “male” and “female,” and the variable “occupation” has the attributes of “student,” “taxi driver,” or “researcher.” Each piece of data has its own attributes and variables; therefore, the investigation of variables will help elucidate the structure of data. In this report, we describe the universal features of the data with variables, addressing the following research question.

Research Question #1: Which types of variables exist in the population of data? What is the distribution of data constructed by certain variables?

Combining different data sources to answer research and business questions is more effective in reducing biases from erroneously measured variables than using a single data source [30], [31]. Although there are technical difficulties in linking heterogeneous data [32], [33], variables can overcome this difficulty and effectively combine data. From the perspective of data exchange and management on platforms, in this study, data were assumed to be the nodes and the variables to be the links between different datasets. By assuming the interactions of data on the platforms to be the network, we elucidate the structural characteristics and relationships of data. As the platforms are not only for shareable data but also for sensitive data, we included data from individuals and private companies and addressed the different behaviors of data in networks based on the sharing conditions. Under these assumptions, we investigated the following question.

Research Question #2: What are the structural characteristics of data networks? What are the types of structural differences between data networks according to their sharing conditions?

The emergence of data platform services has led to intense competition among data platformers. Based on the distributions of data and variables and the structural characteristics of data networks including the sharing conditions, we explored the following question.
Research Question #3: What types of data are tactically important to include on platforms? Which actions are essential for platformers to activate their platforms?

B. DATASET

To answer the three research questions, we examined the structural characteristics of data using data jackets (DJs). A DJ is a framework for summarizing data information while keeping the data confidential [34]. The summarized information of data includes an explanatory text of the data, variable names, format, data sharing conditions, and so on. In other words, a DJ is data of data, i.e., metadata. By using DJs, one can understand the types of data that exist on different platforms and the information included within the data, even if the contents of the data cannot be made public. By consistently converting data of different formats, variable names, and granularities into metadata through common description rules, data in different forms can be uniformly handled.

The main advantage of using DJs in this study was the attributes of variable labels (VLs) and sharing conditions. A VL is an explanatory text about the variable attributes most representative of the data, e.g., “latitude,” “longitude,” “name,” or “age.” In this study, datasets were combined by integrating the variables included in them. Although there are extensive metadata and many databases for specific fields, there are virtually no databases of metadata with uniform description formats across different domains. For example, the metadata provided by data.gov\(^1\) can access the public data of several fields. However, it can only access open data and does not have variable descriptions to facilitate the assessment of data connections. Furthermore, sharing conditions—the conditions for data holders to exchange data with, or provide data to, other parties—are among the key attributes of data, because data are an exchangeable resource on data platforms. Although many data exchange platforms such as The Humanitarian Data Exchange\(^2\) handle data from various fields, they only handle publicly accessed data, while potentially useful data of private companies or individuals are withheld. The description of sharing conditions makes it possible to use both shareable and sensitive data as analysis subjects.

The subjects of this study were the 1,316 DJs that were collected between September 2013 and July 2018 from DJ sites\(^3\) populated by businesspersons, researchers, and data scientists in various fields. We used 3 of the 12 items included in the DJs—namely “name of data,” “sharing conditions,” and “VLs”—in this study. Table 1 shows the basic information about the DJs and VLs. All 1,316 DJs possess at least one VL. We collected DJs in both Japanese and English. Note that, in this paper, the results obtained for the Japanese data have been translated into English. In the experiments, we considered a DJ to be an element representing a dataset and a VL to be an element representing a variable in the data.

Table 2 shows the classification and breakdown of the data sharing conditions. Approximately half the data presented here are shareable, such as the open data made public by government or public institutions. The remaining half are commercial data or generally non-shareable data, for which conditions such as negotiations are required. Accordingly, to simplify the comparison based on sharing conditions for this study, we assessed the data by classifying them into shareable data and sensitive data. “Generally sharable” data were considered to be shareable data. Similarly, we classified data that were “not generally shareable” as sensitive data. Since data holders enter only permissible information into the DJs, the sharing conditions were not included in several DJs. Consequently, data whose sharing conditions were unknown were not employed in the comparison.

Figure 1 illustrates our approach and provides an example of the DJs. First, data holders prepare their data and provide information on those data as DJs to place them on the platform (Fig. 1(a)). We extract the “name of data,” “sharing conditions,” and “VLs” description items from those DJs (Fig. 1(b)). We then analyze the structural characteristics of the data and variables according to the sharing conditions (Fig. 1(c)). By connecting each DJ through the VLs, we create the data network. Under the assumption that the data platform

---

1. https://www.data.gov/
2. https://data.humdata.org/
3. https://datajacket.org/?lang=english
is the data network, we analyze the behaviors and the characteristics of data using the network approach (Fig. 1(d)).

III. STRUCTURAL CHARACTERISTICS OF DATA AND VARIABLES

A. FREQUENCY OF VARIABLES IN DATA

Figure 2 shows the top 10 variables by their frequency of appearance in all 1,316 data, the 654 shareable data, and the 571 sensitive data. The variables indicating location, such as “latitude,” “longitude,” and “address,” and those indicating time-series information, such as “year,” “date and time,” and “time,” appear most frequently. However, 87.5% of the total variables appear once (5,771 variables), and 8.0% appear twice (528 variables). In other words, variables that appear once or twice account for 95.6% of the total number. Although variables such as “latitude” and “longitude” appear as frequently as 80 times, they account for a very small proportion of the total. The mean and median for the frequency of appearance of all variables were 1.37 and 1.00 times, respectively. Variables with a high frequency of appearance constitute less than 1% of the total frequency.

To understand the distribution of variables, we defined the probability of a variable appearing m times as \( p(m) \). However, \( p(m) \) is small when \( m \) is large, and accordingly there are very few \( m \) for which \( p(m) > 0 \). Consequently, the log–log graph of the distribution of the appearance frequency has a weak trend toward noise. Figures 3(a–c) are the rank–frequency plots for each type of data, which are equivalent to the cumulative distribution functions [35]. The distribution of variables is in accordance with the relation, \( p(m) \propto m^{-\gamma} \), which becomes a power distribution curve where the power-law index, \( \gamma \), is 2.30. While several events in society obey Gaussian distributions, other phenomena follow power-law distributions, such as Internet topology [36] and the timing and magnitude of earthquakes [37]. \( \gamma \) is derived from the slope of distribution in the log–log graph. In the case of data variables, \( \gamma \) represents the degree of concentration on variables with low frequencies of appearance. The results show that the data on the platforms do not possess many variables with high frequencies of appearance, but rather an assortment of numerous variables with low individual frequencies of appearance.

B. DISTRIBUTION OF VARIABLES IN DATA

The overall dataset holds 6.86 variables on average, with 5.00 variables as the median (Table 1). In addition, as evidenced by the fact that the maximum number of variables held was 119 and the minimum number was 1, the number of variables held by each dataset varies significantly with data. Data with an extremely large number of variables were large-scale data acquired by sensors, such as healthcare-related data.
Further examination of the data revealed that 195 data have one or two variables (approximately 15% of the total). Meanwhile, approximately 80% of the data have 3–15 variables. Thus, approximately 95% of the total dataset was accounted for by data with 1–15 variables. Designating the proportion of data with \( l \) variables as \( p(l) \), Fig. 3(d) presents a rank-frequency plot of the number of holding variable distributions for each data. Except for data with less than three variables, the distribution is in accordance with \( p(l) \propto l^{-\gamma} \), with \( \gamma = 2.96 \). These results further support that the population of data is not composed of data with many high-frequency variables, but a small number of diverse variables.

**C. DIFFERENCES ACCORDING TO DATA SHARING CONDITIONS**

Although the total number of data is somewhat greater in the case of shareable data, the number of types of variables is approximately equal for both shareable and sensitive data (Table 1). The distribution of variables is in accordance with \( p(m) \propto m^{-\gamma} \), where \( \gamma \) is 2.26 for the shareable data and 2.54 for the sensitive data (Figs. 3(b) and (c)). For a smaller \( \gamma \), data with a smaller number of variables form a large proportion of the population. Thus, shareable data have more common variables, while sensitive data have more data-specific variables.

The mean number of variables for each type of data is somewhat higher in sensitive data, but the median is 5.00 for both types of data. The distribution of data with three or more variables is in accordance with \( p(l) \propto l^{-\gamma} \) for both the shareable and sensitive data where \( \gamma = 3.13 \) for the shareable data and \( \gamma = 2.64 \) for the sensitive data (Figs. 3(e) and (f)). Thus, shareable data are composed of data that do not have many variables. The upper limit of variables in some shareable data is 50. In contrast, the fact that \( \gamma \) is lower for sensitive data than for sharable data means that data with extremely large numbers of variables appear more frequently in sensitive data.

While shareable and sensitive data largely exhibit similar characteristics, the difference between the two types appears in the types of variables present (Figs. 2(b) and (c)). In the case of shareable data, open administration variables such as “latitude,” “longitude,” “year,” and “address,” appeared frequently. In contrast, in the case of sensitive data, the frequencies of variables acquired by the Internet of Things (IoT) sensors and the variables derived from personal information, such as “sex,” “date and time,” and “time” and “age,” appeared frequently. Interestingly, only 207 types of variables are common to shareable data (3,198 types of variables in total) and sensitive data (3,206 types of variables in total). Thus, the types of variables acquired as data significantly differ between shareable and sensitive data. In short, although different types of variables can be obtained and stored as data in different ways, the distributions and structures of data have some similarities, such as the number of types of variables.

**D. SUMMARY OF RESEARCH QUESTION #1**

In this section, we quantitatively discussed the structural characteristics of data on platforms based on the distributions of data and variables. The two major findings described in this section can be summarized as follows. First, despite the absence of a unified data design, a set of orderly features were observed in the power distribution. Second, differences appeared in the types and structures of the data variables according to the sharing condition.

The fact that the number of variables in the data varies proportionally by a power law suggests that the same mechanism drives both small- and large-scale data regardless of the storage format, type, and size of data. Particularly, although data were collected from vastly different areas and sources, a typical power distribution was observed, implying a rule behind the variety of data. Additionally, the data population following this rule was composed of completely different, diverse variables. The experiment revealed that the population was not composed of many high-frequency variables, but
a wide variety of low-frequency variables. Furthermore, the population was largely composed of data with few variables. The attribute each type of data has at the micro-level causes a structural change in the distribution at the macro-level. Taken together, this implies that there is a universal mechanism underlying the different data.

Although shareable data and sensitive data have similar distributions at the macro level, they have some shared condition-specific characteristics. First, in the frequency distribution of variables, \(\gamma\) for the sensitive data was higher than for the shareable data. Thus, sensitive data are largely composed of variables unique to each type of data, which may be due to a lack of communication at the time of data design. Some shareable data, such as papers and government data, are designed uniformly, and many data have been published. Thus, the holders of shareable data can learn the design drawings of other data. In fact, shareable data were created to link them with other data, as is the case with Linked Open Data and other open data whose schema is standardized.

On the one hand, as sensitive data were not assumed to be public or shared with others, no design drawings of data or the types of variables acquired as data are available. Therefore, most variables have low frequencies and are unique to the data, and the number of variables in each type of data is greater—ranging from 1 to 119—compared with sharable data. This increases the \(\gamma\) value for the number of variables in the sensitive data and decreases \(\gamma\) for the distribution of the number of variables in the sensitive data. On the other hand, \(\gamma\) is large for the distribution of the number of variables in the shareable data, and even the data having “many variables” only have approximately 50 variables at most. In other words, many shareable data have the same number of variables.

Based on the aforementioned two findings, although there are differences in the variable type at the micro-level and some differences in the variable frequency distribution and the distribution of variables in the data at the macro-level, there are no overall significant differences between data of different sharing conditions. This suggests that, despite there being difference in the variables acquired, there was a degree of commonality in the process of designing and generating the data. By assuming the data to be world events recognized and described by various people with variables and attributes [38], the similarity of data structure represents the commonality of the frames through which people observe the world. In other words, the number of variables in data is the range of variables that people can recognize and describe. The universal mechanism that drives all data has not been discussed in the literature, and hence these findings are among the contributions of this study to existing research.

IV. ANALYSIS OF DATA NETWORKS

A. STRUCTURAL CHARACTERISTICS

The combination of datasets is achieved by integrating the variables included in the datasets. For example, let us assume that the variables “time” and “date” are common to the

| TABLE 3. Characteristic values of the data networks. |
|-----------------------------------------------|
|                                | Largest component | Shareable data | Sensitive data |
|-----------------------------------------------|
| Number of nodes (|N|)                          | 798             | 434            | 307            |
| Number of links (|E|)                          | 16,076          | 6,663          | 2,460          |
| Average degree (|E|)                          | 40.3            | 30.7           | 16.0           |
| Density (\(\rho\))                        | 0.051            | 0.071          | 0.052          |
| Cluster coefficient (\(C\))                 | 0.702            | 0.703          | 0.648          |
| Assortativity (\(r\))                      | 0.489            | 0.568          | 0.442          |
| Average shortest path length                 | 3.26             | ---            | ---            |

FIGURE 4. (a) Largest components of the data network. (b) Example of the tendency of shareable data to connect with each other and create clusters. (c) Example of sensitive data existing to fill the spaces between shareable data.

“sales data of supermarkets” and “daily weather data.” By combining these data based on their common variables, a data user may learn about the relationship between, for example, sales of beer and changes in temperature. In other words, the combinability of data depends on whether the data involved share common variables. In this section, under the assumption that the data platform is the data network, we analyze the behaviors and characteristics of data on a platform. To express this model, the data network that mediates the variables is represented by an undirected graph, \(G := (N, E)\). Here, \(N\) is the set of nodes composed of data \((N = \{d_i|d_i \in D\})\), \(E\) is the set of edges \((E = \{e_{ij}\})\), and \(D\) represents the set of data \((d_i \in D)\). An edge is established when the same variables appear simultaneously in a pair of data, and it is represented by \(e_{ij} = \{d_i, d_j\} \text{ iff } d_i, d_j \in D, d_i \neq d_j\).

Table 3 summarizes the largest component of the data network, and Fig. 4(a) visualizes it. The discussion in this paper focuses on the largest component, because the overall network was divided into 442 components, and the network excluding the largest component was composed of \(\leq 14\) pieces of data. The density of the network is the ratio of the actual number of links to the number of all possible links.
with nodes in the network. When the number of nodes (\(|N|\)) and the number of links (\(|E|\)) are given, the average degree \(\langle k \rangle\) can be calculated by \(\langle k \rangle = 2|E|/|N|\) and the density \(\rho\) by \(\rho = \langle k \rangle/(|N| - 1)\). Additionally, the cluster coefficient describes the proportion of links that are present between neighboring nodes of a certain node. Assuming that \(k_i\) is the number of links possessed by the \(i\)th node and \(L_i\) the number of links present between the neighboring nodes of the \(i\)th node, the cluster coefficient \((C)\) can be expressed as shown in (1) [39]. For our data network, the density \((\rho)\) was low at 0.051, while the clustering coefficient \((C)\) was high at 0.702. Note that the density represents the global density, and the clustering coefficient indicates the local density.

\[
C = \frac{1}{|N|} \sum_{i=1}^{|N|} \left( \frac{2L_i}{k_i(k_i-1)} \right)
\]

The assortativity \((r)\) indicates the degree of correlation between two neighboring nodes \((-1 \leq r \leq 1)\). If we assume that \(M\) is the number of links, and \(k_p\) and \(k_q\) indicate the degrees of nodes \(p\) and \(q\), respectively, combined by link \(s\), the assortativity \((\text{the degree correlation coefficient})\) is expressed as follows [40].

\[
r = \frac{M^{-1} \sum_s k_p k_q - \left( M^{-1} \sum_s \frac{1}{2}(k_p + k_q) \right)^2}{M^{-1} \sum_s \frac{1}{2}(k_p^2 + k_q^2) - \left( M^{-1} \sum_s \frac{1}{2}(k_p + k_q) \right)^2}
\]

When \(r > 0\), high-degree nodes tend to connect with other high-degree nodes. Conversely, when \(r < 0\), high-degree nodes tend to connect with low-degree nodes. The assortativity values of biological system networks such as proteins or the food chain, and engineering system networks such as the Internet tend to be negative [41]. On the other hand, the assortativity values of networks indicating human relationships, such as the co-author relationships for articles or performance relationships for films, tend to be positive [42], [43]. The data network in this case exhibited \(r = 0.489\), which is a high positive value. This indicates that the network structure consists of high-degree data connected to other high-degree data. Moreover, the average shortest path length was 3.26, which is shorter than that of the natural world (6.80 for the protein interactions) or engineering systems (18.99 for the power grid) [39], [41]. These characteristics suggest that the data network resembles the network of human relationships. For comparison, for a co-authorship network, the clustering coefficient, average shortest path length, and assortativity are 0.43, 5.9, and 0.36, respectively [44], and for a network of film actors, they are 0.79, 3.65, and 0.208, respectively [45]. The clustering coefficients are high, the average shortest path lengths are relatively short, and the assortativity is positive and relatively high in those networks. In data, links are provided between pairs that have the same variables. Consequently, data that have similar variables tend to create clusters, and the combinatorial becomes high. Owing to these characteristics, the data network was locally dense but globally sparse.

When the data network was separated according to sharing conditions, the characteristics of the networks were different. The average degree of the network of shareable data was high at 30.71, and the density was high at 0.071, compared with the corresponding values for the network consisting only of sensitive data for which the average degree was 16.05 and the density was 0.052. In addition, the number of links between the pairs of sensitive data was 2,460, whereas the number of links between the pairs of shareable data was 6,663, which is approximately 2.7 times larger than the former. Since pairs of shareable data have many variables in common, as discussed in the previous section, the shareable data become highly interconnected and tend to form clusters (Fig. 4(b)). Conversely, since the pairs of sensitive data have fewer variables in common, they have a low likelihood of connecting and constitute a sparse network. However, sensitive data seem to exist to fill the spaces between shareable data in Fig. 4(c). We found that the number of links between the pairs of sensitive and shareable data was 5,146, which is approximately 2.1 times larger than the number of links between the pairs of sensitive data. Thus, there are more shareable–sensitive data links than sensitive–sensitive data links.

### B. CENTRALITY OF THE DATA NETWORK

Table 4 shows the top 10 data in terms of degree centrality—the extent to which a certain node is connected with another node. When \(k_i\) is the number of links possessed by the \(i\)th node, the degree centrality of the \(i\)th node can be calculated by (3).

\[
C_D (i) = \frac{k_i}{|N| - 1}
\]

The higher the value, the greater the data linkage and the higher the centrality in the network. We used this measure to elucidate which types of data play a central role in the data platforms. For example, “parking usage data” included “latitude,” “longitude,” “address,” and “time,” whereas “Facebook data” contained “latitude,” “longitude,” “location,” “time,” and “friend data,” which have high frequencies of appearance among the variables. “Latitude” and “longitude,” the variables chiefly associated with locational information, were also included in the dataset “public toilet information.” Moreover, eight of the top 10 data with degree centrality \(\leq 0.160\) contained “latitude” and “longitude.” Thus, although the data network is globally sparse, data with variables indicating the location, such as “latitude,” “longitude,” and “address,” could act as hubs.

In contrast, Table 5 shows the top 10 data in terms of betweenness centrality—the extent to which a certain node connects different groups. The betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. The betweenness centrality of the \(i\)th node is calculated by (4), where \(\sigma_{st}\) is the
TABLE 4. Top 10 data for degree centrality.

| Data name                        | Degree centrality ($C_B$) |
|----------------------------------|---------------------------|
| Parking usage data               | 0.216                     |
| Facebook data                    | 0.178                     |
| Public toilet data               | 0.174                     |
| Twitter data                     | 0.173                     |
| Data of online healthcare shopping| 0.173                     |
| Epicenter list                   | 0.171                     |
| Ozone layer data                 | 0.168                     |
| Data about ultraviolet rays      | 0.168                     |
| GPS data                         | 0.168                     |
| Data of personnel affairs        | 0.167                     |

Note: Bold font indicates sensitive data types.

The number of shortest paths of the $s$th and $r$th nodes, and $\sigma_{sr}(i)$ is the number of those paths passing through the $i$th node.

$$C_B(i) = \sum_{s,t \in N} \frac{\sigma_{sr}(i)}{\sigma_{st}}$$  (4)

The higher the betweenness centrality value, the more the data functions as a bridge in the network. We used this measure to determine which types of data connect various data from different fields. For example, “transportation data” contained “time” and “traffic volume,” which have high frequencies of appearance among the variables. “Latitude,” “longitude,” “country,” and “personal ID” were included in “World happiness data.” In addition, “date and time,” “time,” “personal ID,” etc. were included in “text data of social networking services (SNSes).” There is a tendency for the centralities of location data, personal data, and time-series data to be relatively high. Location data are likely to combine with other data in various contexts. In addition, personal data are extracted from various aspects of human activity for certain purposes. If these data share the same personal IDs, their combinability is high. Furthermore, the variable representing time connects various data in different fields. Data including such variables have high combinability not only with similar data, but also (potentially) with data from other areas.

We also compared the centrality values according to the sharing conditions. In Tables 4 and 5, the types of data in bold are sensitive data. Sensitive data have the potential of being hubs, given that six of the top 10 data for both centrality indices are sensitive data. Furthermore, the average degree centrality of the sensitive data is 0.0447, while that of the shareable data is 0.0558. Thus, although data with an extremely high degree centrality exist among the sensitive data, shareable data generally have a higher degree centrality in the network. On the other hand, although the network of sensitive data has a lower average degree and was sparser than the network of shareable data, the average betweenness centrality of sensitive data is 0.0036, i.e., 1.63 times higher than that of shareable data. Furthermore, in the previous subsection, we discussed that more sensitive data exist between shareable data than between other sensitive data. Thus, the results show that sensitive data may play the role of connecting the data of different fields.

C. DEGREE DISTRIBUTIONS

The connections between data according to the degree distribution of the data network are discussed here. Figure 5(a) shows a rank–frequency plot when the proportion of data with degree $k$ is $p(k)$. Nodes with lower degrees decay gradually; however, the decline becomes abrupt when the degree reaches approximately 60. The rank–frequency plot is in accordance with the exponential function, $p(k) = 6.60 \times 10^2 \exp(-3.00 \times 10^{-2})$. The characteristics of this degree distribution imply that data cannot be linked with other data limitlessly. Most data have degrees between 1 and 100, and data with degrees in excess of 100 are rare. In other words, there are no “all-purpose data” that can connect with various types of data, and there is a limit to the number of connections that one piece of data can have. Accordingly, clusters are formed around data that locally perform as hubs, resulting in a locally dense structure.

Next, we compared the degree distribution of the data network according to the sharing conditions (Figs. 5(b) and (c)). In both shareable and sensitive data, the characteristics of an exponential distribution are observed, where the nodes with low degrees form a gradually declining straight line. However, this distribution declines abruptly when the degree
FIGURE 5. Degree distributions of the data networks (log–log graphs of the rank-frequency plots).

becomes high. The rank–frequency plot of the degree distribution is in accordance with the exponential distribution, \( p(k) = a \exp(-bk) \), where \( a = 4.64 \times 10^2 \), \( b = 2.40 \times 10^{-2} \) for the shareable data and \( a = 2.89 \times 10^2 \), \( b = 2.60 \times 10^{-2} \) for the sensitive data. For the shareable data, the decay in the low-degree portion is gradual, while the decline becomes abrupt when the degree reaches 60. In contrast, in the case of sensitive data, this abrupt decline occurs when the degree reaches 64. These results show that the probability of shareable data having a high degree is greater than that for sensitive data. In other words, although the data network is globally sparse, shareable data have a higher degree and are likely to be locally centralized.

From a global perspective, the data network has an exponential distribution. However, a closer observation reveals that the distribution is a gradually declining straight line in the low-degree region and decays sharply in the high-degree region, which is consistent with a combination of two power distributions (a double power-law distribution [46] and a double Pareto law [47], [48]). Furthermore, the network showed a high, positive assortativity, and consequently the groups of data with low degrees and high degrees have fewer links and each group of data forms a different community. Thus, the data network is divided into two communities—a community consisting of low-degree data and another one consisting of high-degree data—and power distributions appear in both communities. Furthermore, data with a high number of links are relatively rare among the data with degrees less than 60, eventually forming a loose power distribution. In contrast, those data with degrees of 60 or higher are already combined with other high-degree data, but data with many links are infrequent in the community, resulting in a power distribution. In addition, the number of low-degree data was 521, while the number of high-degree data was 277. Thus, the community of low-degree data is larger. This result shows that not only were there sparse areas in the network, which accounted for two-thirds of the network, but also dense communities that obeyed a power distribution.

D. SUMMARY OF RESEARCH QUESTION #2

Three major findings were described in this section. First, we applied the complex network approach to understand the behaviors of the data connections on the platform and found that the network was locally dense but globally sparse, similar to a human relationship network. Second, the data network has the characteristics of an exponential distribution globally, but it is possible that the distribution is a double power-law distribution. Third, the data networks corresponding to the different sharing conditions express different characteristics. The degree distribution of a random network is approximated by the Poisson distribution and the hubs are absent from a random network owing to the uniformity [39]. The fact that our data network obeys an exponential distribution implies that non-uniformity appears in the connections between the data and hubs in the network. Because data networks are created by the commonality of variables, similar pairs of data tend to be easily connected. Furthermore, because of the positive high assortativity, the high-degree data are closely connected to each other and the low-degree data tend to link each other. Consequently, clusters are formed around the data behaving like hubs, resulting in a locally dense structure.

Here, the data network was interestingly observed to have a structure more like the network of human relationships than those of the natural world or engineering systems, considering the characteristics and degree distribution of the data network. The degree distribution of the network of human relations through mobile communications is not a power distribution typified by a scale-free network, but rather an exponential distribution [49]. The exponential distribution is caused by the time limitations of people and the strength of human relationships. That is, very few people are close friends with many other people, and they have almost no time to call all their friends in their busy lives. Since there are no central figures who behave as hubs for connecting with various other people, the network has many small local
groups with strong connections that can be called “close friends.” At the same time, even in the same human relationship network, the degree distributions of SNSes, such as Twitter, are power distributions [50]. Unlike in mobile communication networks, in these networks it is possible to expand friendships without limitations, and there are almost no restrictions on continuing friendships.

In data, the limitation of human cognition may result in human network-like characteristics. The data consist of variables recognized and used by humans. Although there are types of data that consist of an enormous number of variables such as Big Data, the data are not always generated according to a unified or shared design drawing, and the probability of the occurrence of data with many variables is low. The population of data is therefore composed of data with small numbers of a variety of variables. Thus, even if the data pairs share some variables, not all data are aware of each other’s variables, resulting in a globally sparse network. Furthermore, because data acquired with similar intentions and purposes of use have links, as with human-related networks, data clusters with similar variables (which can be called “best friends”) appear locally in the network. Therefore, the data network is globally sparse and locally dense, and obeys an exponential distribution akin to human-related networks. As it is almost impossible to make friends with all people, it can be said that data cannot have limitless links with other data. That is, “all-purpose data” that can connect with all types of data do not exist, but hub data and clusters do appear among similar data.

Secondly, we found that the distribution appears exponential when viewed globally, but the data network distribution consists of two power distributions when viewed closely. This pattern of a separated power distribution also appears in the distribution of human communities [50], [51]. To clarify the community generation mechanisms of data networks, it is necessary to create a model using a micro-level approach based on the type and distribution of variables linking the network.

Thirdly, the number of links between sensitive data and shareable data is greater than that between sensitive data alone. The average degree centrality of shareable data is higher, while the average betweenness centrality of sensitive data tends to be higher. As discussed in Section III, the holders of shareable data can learn the design drawings of other data, while sensitive data are created and maintained privately, resulting in a non-standard creation.

A new question then arises: Why sensitive data are present between shareable data and how do they connect different areas? The answer is that sensitive data were acquired in order to understand events that cannot be explained by shareable data alone. As discussed in the previous section, the types of variables in shareable data and sensitive data are entirely different. Thus, sensitive data cover different areas from shareable data and eventually appear at positions connecting shareable data (Fig. 4(c)). In the DJs, there is an item providing “comments (supplementary information about the data).” For example, looking at “comment data of an Amazon video,” which was registered as shareable data, there was the comment, “SNS [sensitive] data must be combined in order to obtain new knowledge.” In addition, there was the comment, “We need the toilet data of other stations;” to call for sensitive data in the shareable data type “toilet data of Tokyo Station.” Thus, sensitive data appear to connect shareable data to verify the hypotheses that cannot be proven using shareable data alone. Based on the above discussion, it is concluded that shareable data can easily form clusters with each other in the data network, and sensitive data appear at positions connecting these clusters.

V. DISCUSSION

A. TACTICS FOR DATA PLATFORMERS

Based on the results discussed in the previous sections, we have identified some potential tactics for data platformers to promote data exchange among various domains. This section addresses Research Question #3. Firstly, the fact that the data have numerous variables does not mean they have extensive connections with other data. The correlation coefficient between the number of variables in each type of data and the network degree was 0.137. Furthermore, according to the variable distribution, most data types are composed of small number of variables. Thus, data platformers do not necessarily have to prepare data with many variables.

Secondly, it may be beneficial to upload several sets of sensitive data on the platform, as they have high average betweenness centrality and are valuable for connecting data from different areas. In addition, the utility value—the expectation for data utilization and the frequency of browsing on a website—is significantly higher for sensitive data [52]. However, as a network with only sensitive data is sparse, shareable data must be connected to the sensitive data.

Thirdly, we found that the network was highly assortative, locally dense, and globally sparse. Therefore, introducing data that would confer lower associativity and higher density to the platform may be effective in preventing the “silod” condition that tends to create strong local connections (similar to human relationships [53], [54]) and facilitating linkage with data from other domains. To promote integration of cross-disciplinary data, if trends toward high assortativity and low link density are observed in the platform, it may be beneficial for platformers to provide data that increase the betweenness centrality.

Finally, sensitive data and shareable data are composed of completely different types of variables. Considering the possibility of data integration and combination, it will be desirable for platformers to share information about the distribution and types of variables on the platform with the data holders.

B. LIMITATIONS AND FUTURE WORK

The key contributions of this study include an analysis of the characteristics of variables and the macro-level data networks
from empirical data, and an examination of the potential tactics for data platformers. However, many issues need to be explored in the future. Firstly, we introduced DJs because they cover many fields of data, but it may have introduced data selection bias. It is important to enhance the validity of these results by using datasets from other data platforms.

Secondly, in this study, we investigated the behaviors of the data on the platform using the network approach by assuming that the data are related to each other and have connectivity. However, there are many other factors to consider, such as the value of data, marketability, reliability, and privacy. Whether the network approach is optimal for data platforms can be debated. Thus, other approaches used in SNSes or agent interaction [55], [56] for improving the understanding of a data platform should be considered in the future. If a data platform is the place where people meet via data, the network approach is appropriate. The challenge, however, is network dynamics, since few static networks—which are approximations of dynamic networks—exist in the real world. According to a study by Barabási and Albert [57], various models have been proposed to explain the dynamic events in global networks [58], [59]. It is therefore necessary to construct a growth model to evaluate how the introduction of new data would change the relationships of older data on the platforms, in consideration of the “winner-takes-all” phenomenon [60].

Thirdly, there is the possibility that the data network is a crossover distribution represented by power laws with exponential cutoffs [61]. This study did not focus on determining the distribution that would best fit the dataset. For the scope of this research, it was only important that the distribution of data had a long tail. In other words, it was important that there were no exceptional or typical data, or extreme inequality between the data populations. The fact that the frequency follows a long-tailed distribution regardless of the population size suggests that the same mechanism is driving both small- and large-scale data. To improve the understanding of the mechanism, a data growth model must be established in future work based on the simulations in this study.

Finally, the fourth challenge is the similarity of data. In this study, we assumed that the commonality of variables would link data, but contextual links via data usage are also important in searching methods for data utilization and potential data combinations [62], [63]. A more rigorous evaluation of the similarity of data pairs should be undertaken by treating with continuous values not only variables but also the common context of data.

VI. CONCLUSION

The goal of this study was to deepen our understanding of the structural characteristics of data on a platform and their relationships using a variable-based network approach. The main contributions of this work are a novel framework for observing data platforms and the analysis of the behavioral characteristics of data on them. By observing the characteristics of data and variables, we found that the data network had a structure that was locally dense and globally sparse, which is quite similar to networks of human relationships. Moreover, the differences between shareable data and sensitive data were elucidated, which is important for data platforms in a digital economy. Finally, we discussed potential tactics for data platformers to utilize our findings.

Data platforms are new markets in the digital economy. The network analysis provided some important insights on the behavior of data on platforms. However, it is just one of the possible approaches. Although there are some issues to consider going forward, such as the regulation, quality, and reliability of data, we believe the research presented herewith will contribute to the development of data platforms in the future.

ACKNOWLEDGMENT

The authors would like to thank Editage (www.editage.com) for English language editing.

REFERENCES

[1] J. Manyika, M. Chui, P. Groves, D. Farrell, S. V. Kuikten, and E. A. Doshi, “Open data: Unlocking innovation and performance with liquid information,” McKinsey Global Inst., Washington, DC, USA, Tech. Rep., 2013.
[2] J. Manyika, M. Chui, P. Bisson, J. Woetzel, R. Dobbs, J. Bughin, and D. Aharon, “The Internet of Things: Mapping the value beyond the hype,” McKinsey Global Inst., Washington, DC, USA, Tech. Rep., 2015.
[3] E. Rabinovich and S. Cheon, “Expanding horizons and deepening understanding via the use of secondary data sources,” J. Bus. Logistics, vol. 32, no. 4, pp. 303–316, Dec. 2011.
[4] L. M. Elfram and W. L. Tate, “The use of secondary data in purchasing and supply management (PSM) research,” J. Purchasing Supply Manage., vol. 22, no. 4, pp. 250–254, Dec. 2016.
[5] M. Balazinska, B. Howe, and D. Suciu, “Data markets in the cloud: An opportunity for the data as a community,” Proc. VLDB Endowment, vol. 4, no. 12, pp. 1482–1485, 2011.
[6] Y. Ohsawa, H. Kido, T. Hayashi, C. Liu, and K. Komoda, “Innovators marketplace on data jackets, for valuating, sharing, and synthesizing data,” in Knowledge-Based Information Systems in Practice, vol. 30, New York, NY, USA: Springer, 2015, pp. 83–97.
[7] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, “An online pricing mechanism for mobile crowdsensing data markets,” in Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc), Chennai, India, 2017, pp. 1–10.
[8] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, “A survey on big data market: Pricing, trading and protection,” IEEE Access, vol. 6, pp. 15132–15154, 2018.
[9] H. Mano, “EverySense: An end-to-end IoT market platform,” in Proc. Adjunct Proc. 13th Int. Conf. Mobile Ubiquitous Syst., Comput. Netw. Services (MOBIQUitous), Hiroshima, Japan, 2016, pp. 1–5.
[10] Y. Ejiri, E. Ikeda, and H. Sasaki, “Realization of data exchange and utilization society by blockchain and data jacket: Merit of consortium to accelerate co-creation,” in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), Singapore, Nov. 2018, pp. 180–182.
[11] W. Dai, C. Dai, K.-K.-R. Choo, C. Cui, D. Zou, and H. Jin, “SDTE: A secure blockchain-based data trading ecosystem,” IEEE Trans. Inf. Forensics Security, vol. 15, pp. 725–737, Jul. 2020.
[12] S. Abiteboul and J. Systonovitch, “Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation,” J. Data Inf. Qual., vol. 11, no. 3, Jun. 2019, Art. no. 15.
[13] C. Rusche, “Data economy and antitrust regulation,” Inter economics, vol. 54, no. 2, pp. 114–119, Apr. 2019.
[14] F. Stahl, F. Schomm, and G. Vossen, “The data marketplace survey revisited,” Eur. Res. Center Inf. Syst., Münster, Germany, ERCIS Work. Paper 18, 2014.
M. Spiekermann, “Data marketplaces: Trends and monetisation of data goods,” Interэкономics, vol. 54, no. 4, pp. 208–216, Oct. 2019.

K. Sooksatra, W. Li, B. Mei, A. Alrawais, S. Wang, and J. Yu, “Solving data trading dilemma with asymmetric incomplete information using zero-determinant strategy,” in Proc. Int. Conf. Wireless Algorithm. Syst. Appl., Tianjin, China, 2018, pp. 425–437.

F. Stahl, F. Schollm, and G. Vossen, “Data marketplaces: An emerging species,” in Frontiers in Artificial Intelligence and Applications, Amsterdam, The Netherlands: IOS Press, 2014, pp. 145–158.

C. Quix, A. Chakrabarti, S. Kleff, and J. Pullmann, “Business process modelling for a data exchange platform,” in Proc. 29th Int. Conf. Adv. Inf. Syst. Eng., Essen, Germany, 2017, pp. 153–160.

Y. Shen, B. Guo, Y. Shen, X. Duan, X. Dong, and H. Zhang, “A pricing model for big personal data,” Tsinghua Sci. Technol., vol. 21, no. 5, pp. 482–490, Oct. 2016.

X. Cao, Y. Chen, and K. J. R. Liu, “Data trading with multiple owners, collectors, and users: An iterative auction mechanism,” IEEE Trans. Signal Inf. Process. Netw., vol. 3, no. 2, pp. 268–281, Jun. 2017.

J. Van Tasdel, Digital Rights Management: Protecting and Monetizing Content. Waltham, MA, USA: Focal Press, 2016.

L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, “Information security in big data: Privacy and data mining,” IEEE Access, vol. 2, pp. 1149–1176, 2014.

C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen, “Trading data in good faith: Integrating truthfulness and privacy preservation in data markets,” in Proc. IEEE 13rd Int. Conf. Data Eng. (ICDE), San Diego, CA, USA, Apr. 2017, pp. 223–226.

P. Groth and Y. Gil, “Linked data for network science,” in Proc. 1st Int. Conf. Linked Sist., Bonn, Germany, 2011, pp. 1–12.

C. Guéret, P. Groth, C. Stadler, and J. Lehmenn, “Assessing linked data mappings using network measures,” in The Semantic Web: Research and Applications (Lecture Notes in Computer Science), vol. 7295, E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti, Eds., Berlin, Germany: Springer, 2012, pp. 87–102.

J. Euzenat and P. Shvaiko, Ontology Matching. Berlin, Germany: Springer-Verlag, 2007.

L. Zhao and R. Ichise, “Ontology integration for linked data,” J. Data Semantics, vol. 3, no. 4, pp. 237–254, May 2014.

T. Hayashi and Y. Ohsawa, “Matrix-based method for inferring variable labels using outlines of data in data jackets,” in Proc. 21st Pacific-Asia Conf. Knowl. Discovery Data Mining, Jeju, South Korea, 2017, pp. 696–707.

E. Babie, The Basics of Social Research, 7th ed. Boston, MA, USA: Cengage Learning, 2016.

M. Weiner, Z. Quatelli, A. J. Perkins, J. N. Lewis, and C. M. Callahan, “Limitation of a single clinical data source for measuring Physicians’ performance on quality indicators,” J. Amer. Geriatrics Soc., vol. 54, no. 8, pp. 1256–1260, Aug. 2006.

G. Ridder and R. Moffitt, “The econometrics of data combination,” Handb. of Econometrics, vol. 6, pp. 5469–5547, Jan. 2007.

E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches,” in Contemp. Syst. Eng., vol. 6, pp. 546–554, Jan. 2007.

M. E. J. Newman, Networks. Oxford, U.K.: Oxford Univ. Press, 2010.

M. E. J. Newman and J. Park, “Why social networks are different from other types of networks,” Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., vol. 68, no. 3, Sep. 2003, Art. no. 036122.

S. Redner, “Teasing out the missing links,” Nature, vol. 453, no. 7191, pp. 47–48, Apr. 2008.

M. E. J. Newman, “Cohortship networks and patterns of scientific collaboration,” Proc. Nat. Acad. Sci. USA, vol. 101, no. 1, pp. 5200–5205, Jan. 2004.

D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” Nature, vol. 393, pp. 440–442, Jun. 1998.

S. Jalan, K. Kanhaiya, A. Rai, O. R. Bandapalli, and A. Yadav, “Network topologies decoding cervical cancer,” PLoS ONE, vol. 10, no. 8, 2015, Art. no. e0135183.

J. Van Tassel, The double Pareto-longnormal Distribution—A new parametric model for size distributions, Commun. Statist. Theory Methods, vol. 33, no. 8, pp. 1733–1753, Dec. 2004.

W. Li, Q. A. Wang, L. Nivanen, and A. Le Méhaute, “How to fit the degree distribution of the air network?” Phys. A. Stat. Mech. Appl., vol. 368, no. 1, pp. 262–272, Aug. 2006.

J.-P. Omella, J. Saramäki, J. Hyvönen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabási, “Structure and tie strengths in mobile communication networks,” Proc. Nat. Acad. Sci. USA, vol. 104, no. 18, pp. 7332–7336, Apr. 2007.

W.-Y. Shin, B. C. Singh, J. Cho, and A. M. Everett, “A new understanding of friendships in space: Complex networks meet Twitter,” J. Inf. Sci., vol. 41, no. 6, pp. 751–764, Nov. 2015.

Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, “Analysis of topological characteristics of huge online social networking services,” in Proc. 16th Int. Conf. World Wide Web (WWW), Alberta, CA, Canada, 2007, pp. 835–844.

T. Hayashi and Y. Ohsawa, “Comparison between utility expectation of public and private data in the market of data,” in Proc. 20th Int. Conf. Knowl. Based Intell. Inf. Syst., York, U.K., vol. 96, 2016, pp. 1267–1274.

S. C. Goh, “Managing effective knowledge transfer: An integrative framework and some practice implications,” J. Knowl. Manage., vol. 6, no. 1, pp. 23–30, Mar. 2002.

R. Keast, M. P. Mandell, K. Brown, and G. Woolcock, “Network structures: Working differently and changing expectations,” Public Admin. Rev., vol. 64, no. 3, pp. 363–371, May 2004.

G. Pallá, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” Nature, vol. 435, no. 7043, pp. 814–818, Jun. 2005.

K. Osaka, F. Toriumi, and T. Sugawara, “Effect of direct reciprocity and network structure on continuing prosperity of social networking services,” Comput. Social Netw., vol. 4, no. 1, pp. 1–20, May 2017.

A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” Science, vol. 286, no. 5459, pp. 509–512, Oct. 1999.

J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graphs over time: Densification laws, shrinking diameters and possible explanations,” in Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), Chicago, IL, USA, 2005, pp. 177–187.

P. Holme and J. Saramäki, Temporal Networks (Understanding Complex Systems). Berlin, Germany: Springer-Verlag, 2013.

G. Bianconi and A.-L. Barabási, “Bose-Einstein condensation in complex networks,” Phys. Rev. Lett., vol. 86, no. 24, pp. 5632–5635, Jun. 2001.

A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-law distributions in empirical data,” SIAM Rev., vol. 51, no. 4, pp. 661–703, Nov. 2009.

E. Ruijer, S. Grimmelikhuijsen, M. Hogan, S. Enzerink, A. Ojo, and A. Meijer, “Connecting societal issues, users and data. Scenario-based design of open data platforms,” Government Inf. Quart., vol. 34, no. 3, pp. 470–480, Sep. 2017.

T. Hayashi and Y. Ohsawa, “The difference between variable-based and context-based networks of data using data jackets,” in Proc. 22nd Int. Conf. Knowl. Based Intell. Inf. Eng. Syst., Belgrade, Serbia, vol. 126, 2017, pp. 1740–1747.
TERUAKI HAYASHI received the Ph.D. degree in engineering from The University of Tokyo. He is currently an Assistant Professor of systems innovation affiliated with the School of Engineering, The University of Tokyo, and the Vice Chairman of the Application Committee at the Data Trading Alliance. His specialization is in knowledge structuring for data utilization and scenario creation. He developed the data jacket, a method for summarizing the information in data, as a part of his core research and applied the results internationally to industry, government, and academia. He is also the coauthor of the book Market of Data (Kindaikagakusha, 2017). He was awarded the Dean’s Award by the School of Engineering at The University of Tokyo, in 2017, and the Excellence Award at the 23rd Annual Conference of the Japanese Society of Artificial Intelligence, in 2018.

YUKIO OHSAWA received the Ph.D. degree from the School of Engineering, The University of Tokyo, in 1995. He then worked with the School of Engineering Science, Osaka University (a Research Associate, from 1995 to 1999) and subsequently the Graduate School of Business Sciences, University of Tsukuba, (an Associate Professor, from 1999 to 2005), before moving back to The University of Tokyo. He is currently a Professor of systems innovation with the School of Engineering, The University of Tokyo. In the field of artificial intelligence, he has created a new domain known as chance discovery, which discovers events with significant impact on decision making. He has delivered keynote speeches about chance discovery at conferences such as the International Symposium on Knowledge and Systems Sciences, the International Conference on Rough Sets and Fuzzy Sets, the Joint Conference on Information Sciences, and Knowledge-Based Intelligent Information and Engineering Systems. Chance discovery has been embodied as an innovators’ marketplace; a methodology for innovation that is borrowed from principles of market dynamics. His original concepts and technologies have been published as books and monographs by global publishers such as Springer, Verlag, and Taylor and Francis. His two most important books among these publishers are Chance Discovery (Springer, 2003: E. V Hippe gave the opening) and Innovators Marketplace: Using Games to Activate and Train Innovators (Springer, 2012: L. Leifer gave the opening). He has also published 100 journal articles, and initiated symposia and workshops on data-based approaches to business innovation. He has edited special issues as a guest editor for journals mainly related to chance discovery such as Intelligent Decision Technologies (2016), Information Sciences (2009), New Generation Computing (2003), and the Journal of Contingencies and Crisis Management (2002).

* * *