Expert Systems With Applications 106 (2018) 121–131

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Historical inference based on semi-supervised learning

Dong-gi Lee[a,1], Sangkuk Lee[b,1], Myungjun Kim[a], Hyunjung Shin[a,*]

[a] Department of Industrial Engineering, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon 16499, South Korea
[b] Department of History, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon 16499, South Korea

ARTICLE INFO

Article history:
Received 20 September 2017
Revised 14 December 2017
Accepted 28 March 2018
Available online 5 April 2018

Keywords:
Machine learning
Semi-supervised learning
Historical big data
Genealogy

ABSTRACT

In the past, most historical research has been manually carried out by exploring historical facts reading between the lines of documents. Nowadays, historical big data has become electronically available and advances in machine learning techniques allow us to analyze the vast amount of historical data. From a historical perspective, making inferences about political stances of historical figures is important for grasping historical rivalries and power structures of an era. Thus, in this paper, we propose an approach to the systematic inference of power mechanisms based on a human network constructed from historical data. In this network, humans are linked according to the degree of kinship using genealogy records, and identified by political stances on agendas recorded in the annals of a dynasty as a political force. And then, a machine learning algorithm, semi-supervised learning, classifies humans who cannot identify political stances as political forces that reflect the links of the networks. The data consist of the genealogy of the Andong Gwon clan, a record of family relations of 10,243 people from the 10th to 15th century Korea, and the Annals of the Joseon Dynasty, a historical volume that describes historical facts of the Joseon Dynasty for 472 years and is composed of 1894 fascicles and 888 books. From the data, we construct a human network based on a historically meaningful period (1443–1488), and classify people into two political forces using the proposed method. We suggest that this machine learning approach to historical study could be utilized as a potent reference tool devoid of the subjectivism of human experts in the field of history.

1. Introduction

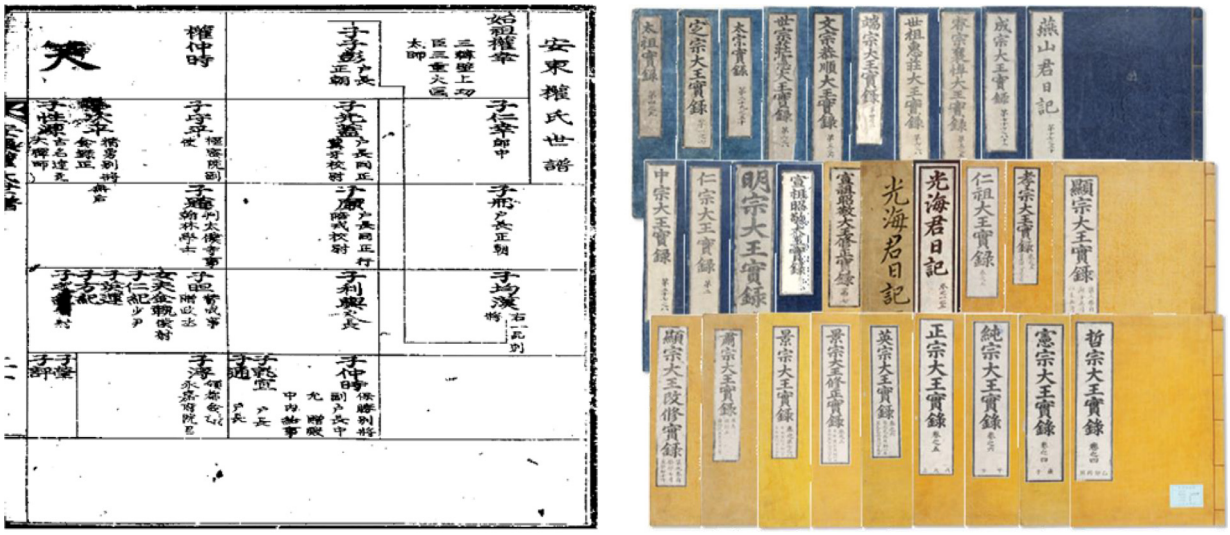
Throughout the recent decade, machine learning has pervaded various fields of study including economics, sociology, political science, natural science, and medical science. These incorporations of machine learning have received much attention due to their successful blend of research and the potential for innovative developments in these fields (Qiu, Wu, Ding, Xu, & Feng, 2016; Sharma, Agrawal, Agarwal, & Sharma, 2013; Singh, Bhatia, & Sangwan, 2007). Nowadays, it is difficult to find an area of study that does not incorporate machine learning. In contrast, it is uncommon to observe the utilization of machine learning in the field of history because of its strong rigidity and the high barriers to entry in the domain. In recent years, however, there has been an increase in the digitalization and availability of historical data (Lee, 2010, 2016a, 2016b; Manabe, 1999; Manoff, 2010). This has begun to lead to initiative attempts at research based on machine learning in the field of history.

Stretching beyond simple descriptive statistics, Bak and Oh (2015) applied a machine learning technique called topic modeling to the annals of a certain dynasty in Korean history in order to determine the inclinations of kings for agendas on various political topics. Whereas, there exists an attempt to visualize historical data using machine learning techniques (Liu, Dai, Wang, Zhou, & Qu, 2017). The authors proposed a visual analytical system for Chinese genealogical data (Lee & Campbell, 2010), which is hierarchical, spatiotemporal, and multidimensional. This system helps improve the feedback from domain experts for the historical literature and helps users understand the structure and contents of Chinese history.

These studies focus on the exploration or description of past historical facts, assisting historians to view historical records from a comprehensive perspective. On the contrary, another objective that is key for research in history is inference on historical facts. In particular, inferring the power mechanism in a specific era helps historian to take a close look at the power structure, that is, the basic principle of implementation of political decision-making in

* Corresponding author.
E-mail address: shin@ajou.ac.kr (H. Shin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.eswa.2018.03.059



(a) (b)

Fig. 1. (a) Screenshot of a scanned version of the AGS, (b) Image of the AJD.

that period. According to the definition of politics, power refers to the ability to influence the behavior of others in the realm of various political, economic, and societal relationships among people (Dahl, 1957). It has the property of producing relative rank or classes among the people who constitute the society. With the relational and interactive nature of power functioning in a society, it is inevitable that individuals will make strategic decisions (Foucault, 1980). We define a power mechanism as the principle or structure of rivalries, political decision patterns, and maintenance of power that acts upon a society or a nation. Such power mechanisms are analyzed and interpreted to understand the overall structure of politics, society, and economy by expressing the power structure or the structure of political choices. Power mechanisms are currently being analyzed intensively in various disciplines, especially in the field of premodern political history, not only for understanding the current political structure, but also for grasping the power structure of past ruling elites. The following section presents additional literature reviews on power mechanisms.

In this study, we analyze the power mechanisms of the Joseon Dynasty by utilizing machine learning on historical data. More specifically, we attempt to classify political forces using existing historical records, the Andong Gwon-ssi Sunghwabo (AGS) and the Annals of the Joseon Dynasty (AJD). The AGS is currently the oldest existing set of genealogy data in Korea and was published in 1476 by a member of the 18th generation of the Andong Gwon clan, Seo Geojung. There are family relation records of 10,243 people over twenty-one generations with names for each individual and their final position at the national bureau. It can be browsed in the Jangseogak Royal Archive (http://yoksa.aks.ac.kr/) as a scanned version, as shown in Fig. 1(a). Later, Lee (2013) digitized it based on the Henry system (Henry, 1935). On the other hand, the AJD is the most renowned historical volume that describes the historical facts of the Joseon Dynasty for 472 years over 25 reigns, from King Taejo (1392) to King Chuljong (1863). It consists of 1894 fascicles and 888 books, comprising 48,646,667 Chinese letters. In 1997, it was registered as a documentary heritage in UNESCO's Memory of the World program. Fig. 1(b) shows an image of a few fascicles of the AJD. The images, complete translation in Korean, and database of the AJD are currently accessible online from http://sillok.history.go.kr.

To analyze the power mechanisms of the era using these historical data, the proposed method begins by constructing a human network. The network consists of nodes, edges, and labels. The nodes indicate people, edges represent blood relations between people in terms of degree of kinship, and labels represent power groups for individuals. For the construction, we used the AGS for the nodes and edges and AJD to determine the labels. To construct the network, we applied a machine learning algorithm called semi-supervised learning (SSL) (Chapelle, Scholkopf, & Zien, 2006; Zhu & Goldberg, 2009; Zhu, Ghahramani, & Lafferty, 2003). SSL is known to be effective when labeled data are limited. People known to be within power groups are very scarce—perhaps only historically important figures—which means labeled data points are limited in the historical data. In such circumstances, SSL is suitable for identifying groups of people who are not defined as belonging to any group. The results inferred by SSL are validated through expert annotation.

The rest of the paper is organized as follows. In Section 2, we further remark on Korean historical data and previous studies on power mechanisms. In Section 3, we briefly review the SSL and its formulation. In Section 4, we present the proposed method, explaining the procedures necessary to process the data for SSL. In Section 5, we explore experiments and the results of applying the proposed method to the human network built from the AGS and AJD. In Section 6, we conclude this paper with some discussion of the insights, novelty, and limitations of the current research.

2. Remarks on Korean historical data and previous studies on power mechanisms

Representative studies on power mechanisms include rivalries between the Powerful Family and New Officials in the late Goryeo Dynasty (Min, 1974; Yi, 1991), Meritorious Elite and Neo Confucian in the early Joseon Dynasty (Lee, 1979; Lee, 1986), faction politics in the late Joseon Dynasty (Chung, 1983; Lee, Lee, Shin, Jung, & Yoo, 1993; Lee, 2000; Lee, 1985), and pro- and con-Japan groups in the Japanese colonial period (Kang, 1980). In the current field of history, however, such studies are mostly based on empirical approaches that interpret and describe historical records by meticulously looking at associated words, phrases, and even meanings between sentences. Although empirical approaches yield enriched interpretations through detailed analysis, they have a disadvantage because of the significant time and effort required to explore the tremendous amount of existing historical data. In recent years, however, there has been an increase in the interest of database creation and extensive efforts to create databases for

large amounts of historical data. Such efforts have yielded the digitalization of historical big data and enabled easier access to these data. Works concerning this trend include standardizing databases for genealogy data, building up a basis for constructing databases related to Korean medieval archeology, and generating a history ontology through web-based data, where it is produced by users. Most notable works regarding Korean history are the History of Goryeo and the AJD, which can be browsed through the Internet. Furthermore, many studies have attempted to collect and scan genealogy data that show blood relations for various families. Starting from the work of Wagner and Song (Lee, 2004; Song & Wagner, 2000), which created a database for the roster of civil examination graduates in the Joseon Dynasty, many historical records have now been digitalized. Extending from the digitalization of historical records, there are growing efforts at interpreting historical facts or phenomenon using digitalized data. Chung and Kim (2011) created a human network oriented on a king in the Three Kingdoms Period based on domestic politics, international relations, family relations, and relations between people (excluding the king himself). By considering the number of connected links in the network, they deduced the governing rules and management methods of each king. Lee and Son (2012) analyzed the periodic patterns of life expectancy and death rates from the 17th to 20th century based on the dates of births and deaths in a genealogy dataset.

3. Semi-supervised learning

In view of the methodology, the inference of power mechanisms can be regarded as a classification of people into two (or more) groups. Here, the indicator for one's membership to one group or another one can be coded as a label $+1$ or -1. However, people known to be within power groups are very rare in the historical data—perhaps only historically important figures. This means it is not practical to use supervised learning algorithms as an inference model. SSL, in contrast, allow us to circumvent this difficulty because it can work with only a few labeled data but at the same time it utilizes a good number of unlabeled data to infer a natural and smooth classification boundary between two rivalry power groups (Chapelle et al., 2006; Kim & Shin, 2013; Wang, Shen, & Pan, 2009; Zhu, 2005). We briefly introduce SSL as follows.

In mathematical notation, the use of SSL is ideal particularly when $n_l \ll n_u$, where n_l is the number of labeled data and $n_u$is the number of unlabeled data. Graph-based SSL is a commonly used algorithm; it utilizes a graph (or a network) to represent the similarities between data points and predicts labels for unlabeled data (Goldberg & Zhu, 2006; He, Carbonell, & Liu, 2007; Shin, Hill, Lisewski, & Park, 2010; Subramanya & Talukdar, 2014). Unlike other machine learning algorithms, such as artificial neural networks (ANNs) (Abraham, 2005; Bishop, 1995), support vector machines (SVMs) (Cristianini & Shawe-Taylor, 2000; Kotsiantis, Zaharakis, & Pintelas, 2006; Schölkopf & Smola, 2002; Shin & Cho, 2007), and decision trees (DTs) (Breiman, Friedman, Stone, & Olshen, 1984; Kass, 1980; Quinlan, 2014), which only use labeled data, SSL can capture the structure of the input space by employing additional unlabeled data in the learning process. Fig. 2 depicts graph-based SSL.

For graph-based SSL, a network $N=(V, E)$ consists of a node set V and an edge set E. For $n(=n_l+n_u)$ data points, V represents a set of data points ($V=\{x_1, x_2, \ldots, x_n\}$) and E represents the similarities between the nodes. In general, the weight w_{ij} of nodes i and j in the similarity matrix W is calculated by the Gaussian function

$$w_{ij} = \begin{cases} exp^{-dist(x_i,x_j)/\sigma^2} & if\ i \sim j, \\ 0\ otherwise \end{cases} \quad (1)$$

where $i \sim j$ indicates that the two nodes are connected, and $dist(x_i,x_j)$ is the distance between x_i and x_j with any distance measure. Weight w_{ij} takes on a value between 0 and 1, where a higher value implies higher similarity between the two nodes. The label set $y=(y_l,y_u)$ consists of labeled nodes $y_l \in \{-1, +1\}(l=1, \ldots, l)$ and unlabeled nodes $y_u=0(u=l+1, \ldots, l+u)$. SSL produces output $\boldsymbol{f}=(f_1,\ldots, f_l,f_{l+1}, \ldots, f_{l+u})^T$ for unlabeled data by minimizing the following quadratic function:

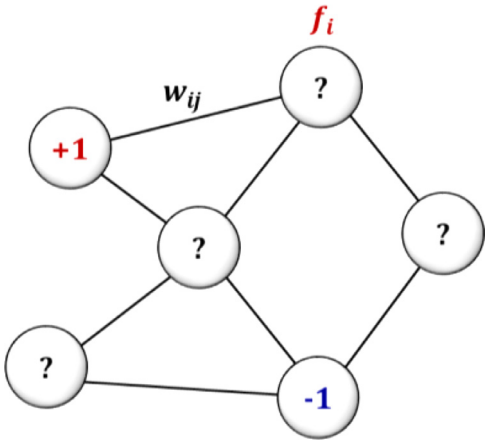


Fig. 2. Graph-based semi-supervised learning (SSL).

$$\min_{\boldsymbol{f}} \ (\boldsymbol{f}-\boldsymbol{y})^T(\boldsymbol{f}-\boldsymbol{y}) + \mu \boldsymbol{f}^T L \boldsymbol{f} \quad (2)$$

where $\boldsymbol{y}=(y_1,\ldots, y_l, 0, \ldots, 0)^T$ and L, the graph's Laplacian matrix, is defined as $L=D-W$, where $D=diag(d_i)$ and $d_i=\sum_j w_{ij}$. The graph-based SSL algorithm has two conditions: (a) the output should be consistent with the initial labeling and (b) the outputs for two connected nodes should not be too different given their similarity. Parameter μ is a user specific learning parameter that controls the two conditions. The solution to minimization problem (2) is given by

$$\boldsymbol{f} = (I+\mu L)^{-1}\boldsymbol{y}, \quad (3)$$

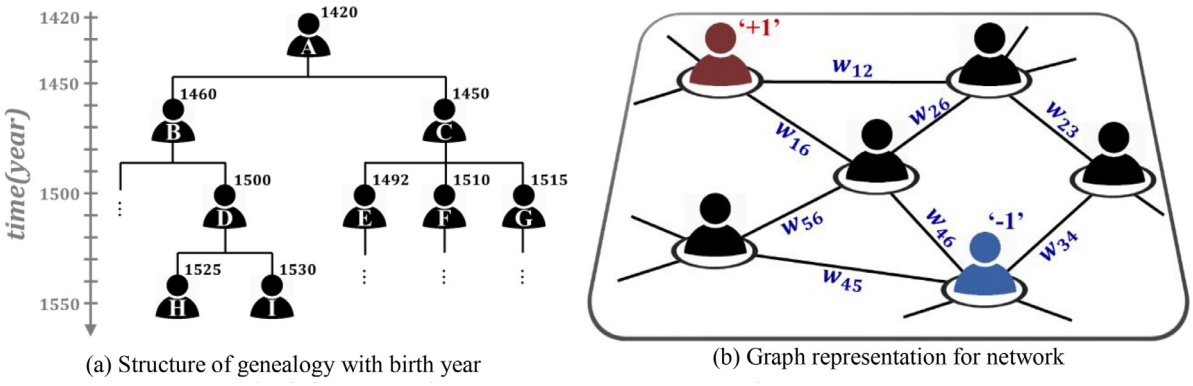
where I is the identity matrix. With output (3), we can classify unlabeled nodes with either -1 or $+1$.

4. Proposed method

To create a network of people in a genealogy, a tree structure representing parent and offspring relationships, as shown in Fig. 3(a), should be transformed into a network structure, as described in Fig. 3(b). In the network, nodes represent people in the Joseon Dynasty, edges represent relationships between people, and labels represent each associated power group. SSL can then be applied to the network to draw inferences on the group memberships of people whose memberships to a certain power group are not known. In general, a network is constructed by representing the data as nodes and similarities between them as edges. Considering the purpose of the present study, however, it is meaningless to represent all the people who are recorded in the genealogy over hundreds of years into one single network. Thus, we constructed a network that only includes people in a contemporary period and the similarities among them. Then, we developed a method for classifying the people into rival power groups based on the information obtained from a series of political agendas in the AJD. Fig. 4 illustrates the process of the proposed method.

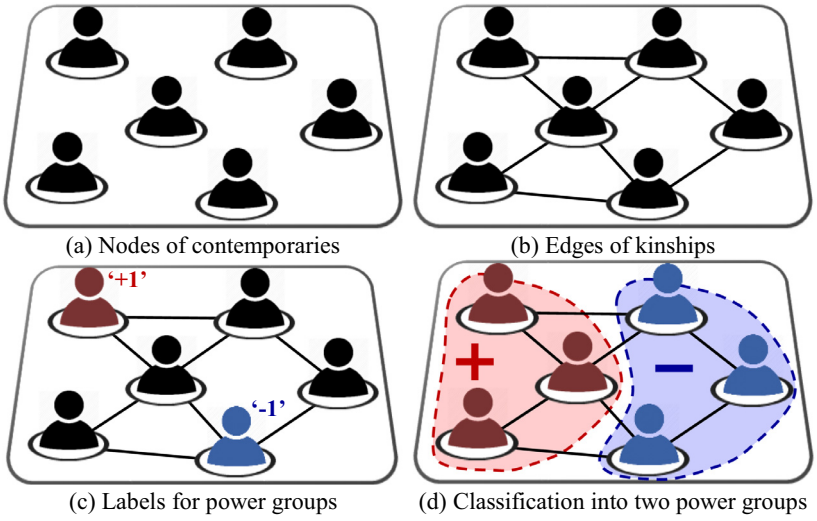
4.1. Network of contemporaries

To construct a network of contemporaries from a genealogy, it is required to define an "orient" (a person with whom we are con-



(a) Structure of genealogy with birth year (b) Graph representation for network

Fig. 3. Structure of genealogy and people network from the genealogy.



(a) Nodes of contemporaries (b) Edges of kinships

(c) Labels for power groups (d) Classification into two power groups

Fig. 4. Process of the proposed method: (a) Select the nodes of the network by defining an orient for contemporaries and the length of the corresponding period. (b) Calculate the edge weights that represent the similarity of the people based on their blood relations. (c) Set the labels of the people based on the information obtained from a series of political agendas in the AJD. (d) Classify people into rival power groups by applying the graph-based SSL algorithm and inferring rivalries for the corresponding period.

cerned) to define a contemporary period and the length of this period. In the simplest method, all the people in the same generation as the orient are added to the network. The depth of the tree specifies the network structure. In such a case, a period of the same depth in the genealogy tree can be stretched to cover more than 100 years. However, it does not make sense to say that people in this network are contemporaries. For instance, a cousin of somebody can be older than his parents; they are same generation, but not contemporaries.

Therefore, we propose a flexible time-window method that sets a historically important person as an orient and uses that person to define the contemporary period. In genealogy data, chronological information such as birth year, death year, and the year of admission to a national bureau are available (but does not include all of the people recorded in the genealogy). Let us denote the person chosen as an orient as a *time marker*. Then, the contemporary period is derived by determining the upper and lower bounds. In addition, we consider that the bounds should be set to cover the valid period over which the time marker exercises his/her power (see Fig. 5). For the upper bound Yr^U, the mean incumbency of people recorded in the genealogy is calculated and added to the time marker's admission year to a national bureau. However, this date cannot exceed the time marker's year of death:

$$Yr^U = min\ \{(YTP_{TM} + INC_{mean}),\ YD_{TM}\}, \tag{4}$$

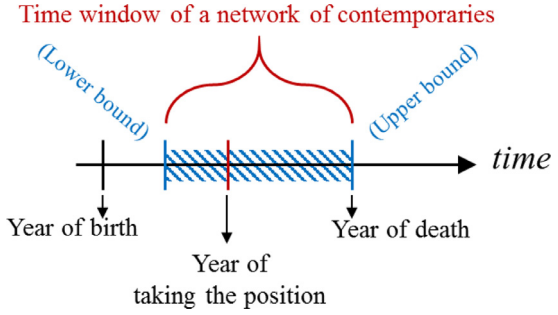


Fig. 5. Concept of the flexible time-window oriented on a time marker.

where YTP_{TM} and YD_{TM} are the time marker's year of taking the position (or admission year to a national bureau) and year of death, respectively. Moreover, INC_{mean} is the average incumbency, calculated from the genealogy. The lower bound Yr^L for the time window similarly set. In our study, a year that is earlier than the time the time marker takes an official position is considered because he/she could have been already influenced by his/her family members even though he/she does not yet have a position. It is calculated using the following equation:

$$Yr^L = \frac{YB_{TM} + YTP_{TM}}{2} \tag{5}$$

where YB_{TM} is the time marker's year of birth. Any person historically who is well-known can be a time marker. Moreover, the length of the contemporary period can vary depending on the time marker. Therefore, this approach provides flexibility when constituting a network from a genealogy.

4.2. Nodes

From the genealogy, people with a record of an admission year to a national bureau within the bounds (4) and (5) are included in the network. Those become nodes of the network. However, because of missing information, the network only contains a small portion of people. Those who do not have time information recorded in the genealogy cannot be included in the network. To overcome such deficiencies, we utilize blood relations to supplement the network with more people: the siblings of the people with time information are added to the network. This incorporates the general conjecture that sisters and brothers are likely to live in a similar period.

4.3. Edges

An edge in the network represents a blood relationship between people in a genealogy and its connection strength is calculated using the degree of kinship. From the tree structure of a genealogy, the degree of kinship is defined as the following: the direct connection of a parent and offspring relationship is a first-degree of kinship, siblings are second-degree kinship, an uncle or aunt is a third-degree of kinship, and this increases in the same manner. A smaller value of the degree of kinship implies closer relations between the two people, therefore the connection strength (or similarity) is larger. We reflect such aspects to calculate the similarity matrix using the following equation:

$$w_{ij} = \frac{2}{1 + e^{k_{ij}}} \tag{6}$$

where k_{ij} is the degree of kinship between person i and j. The range for similarity w_{ij} is $0 \le w_{ij} \le 1$, where $w_{ij} = 1$ for the person him/herself ($k_{ii} = 0$). Moreover, w_{ij} has high values for close relatives and low values for distant relatives. If the value of w_{ij} is below threshold ε, we disconnect the edges (i.e., let $w_{ij} = 0$ for $w_{ij} < \varepsilon$).

4.4. Labels

The labels of the nodes in the network represent memberships of people to a certain power group or the other rival group. However, there are many instances where the records for power information are non-existent or were destroyed for political purposes. Thus, we define two opposing power groups and assign people into one of the power groups based on the AJD, which contains discussions on various agendas and the opinions of officials in the Joseon Dynasty period. More specifically, we compare the opinions of people whose opinions are identifiable in the AJD, for various agendas under the assumption that people in the same power group generally express similar opinions about agendas. Because there can be no perfect consensus or dissension in opinions between two people, we set a quantitative measure for the degree of consensus in opinions using the phi coefficient, which is a measure of association for two binary variables (Boas, 1909; Pearson, 1900; Yule, 1912). Suppose we have the following contingency table for two variables a and b ($a, b \in \{0, 1\}$),

	$b=0$	$b=1$
$a=0$	n_{00}	n_{01}
$a=1$	n_{10}	n_{11}

Table 1
Opinion information of people on five toy agendas.

	Agenda 1	Agenda 2	Agenda 3	Agenda 4	Agenda 5
Person D	Agree	Disagree	Agree	Disagree	Disagree
Person E	N/A	N/A	N/A	N/A	N/A
Person F	Disagree	Agree	Disagree	Agree	Agree
Person G	N/A	N/A	N/A	N/A	N/A
Person H	Agree	Disagree	Agree	Agree	Disagree
Person I	N/A	N/A	N/A	N/A	N/A

Table 2
Phi-coefficient matrix of three people.

	Person D	Person F	Person H
Person D	1.0000	−0.6667	1.0000
Person F		1.0000	−0.6667
Person H			1.0000

where n_{00}, and n_{11} are the observation frequencies for consensus and n_{01} and n_{10} are the observation frequencies for dissension. The phi coefficient φ is calculated with the following equation:

$$\phi = \frac{n_{00}n_{11} - n_{10}n_{01}}{\sqrt{(n_{00} + n_{01})(n_{10} + n_{11})(n_{00} + n_{10})(n_{10} + n_{11})}}. \tag{7}$$

Just as for the Pearson correlation coefficient, φ takes a value between -1 and 1, where values closer to 1 indicate a higher degree of consensus and those closer to -1 indicate a higher degree of dissension. Using the phi coefficient, we define a power group consisting of people with similar opinions in contrast to its rival group. People belonging to same group have the same label and are called labeled people in our study.

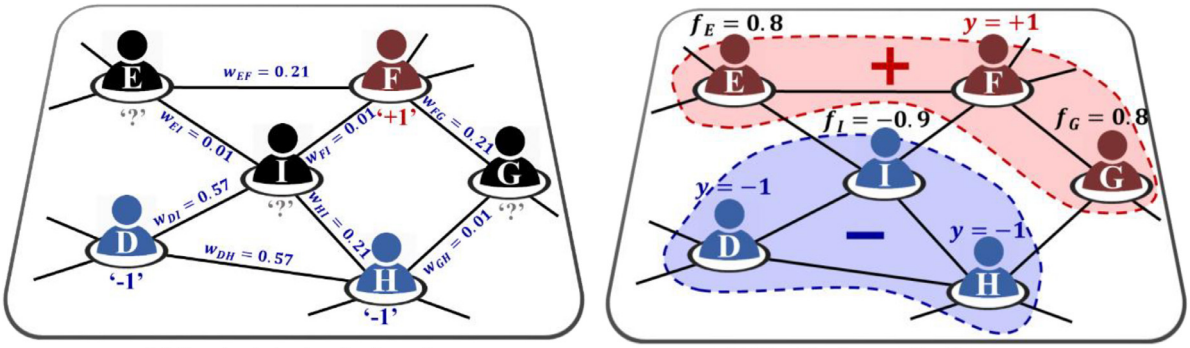
4.5. Classification of power groups

To classify unlabeled people in the network, that is, those who are not identified with any power group via labeling, we use the prediction of a graph-based SSL algorithm. For convenience of understanding, we explain the classification process with a toy example. Suppose we have the genealogy network shown in Fig. 3(a), with person D set as the time marker. Further assume that his birth year, death year, and admission year to a national bureau are 1500, 1560, and 1520, respectively. By applying the flexible time-window, the period between 1510 and 1560 is set from the bounds of the time window. Hence, we can include six people into the network who have records of admission to a national bureau: Persons D, E, F, G, H, and I.

For the edges, we employ the degree of kinship along with (6) to calculate the weights. For instance, Persons D and H (or I) have a first-degree of kinship, which yields 0.57 as the edge weight, and Persons D and E (or F and G) have fourth-degree of kinship, which yields 0.04 as the edge weight.

For labels, we use opinions for five toy agendas, which are shown in Table 1. In this case, there only exists opinion information for Persons D, F, and H, so we apply the phi coefficient only to these three people for labeling. Table 2 shows the resulting values of the phi coefficient for Persons D, F, and H.

With respect to the time marker Person D, Person H has a positive phi coefficient value but Person F has a negative value. Thus, we can define a group that Person D and H belong to, and its rival group, which Person F belongs to. In the notation, we assign label "−1" to the former group's members whereas "+1" is assigned to the latter group's member. For Persons E, G, and I, for whom opinion information does not exist, we assign "0", indicating them as unlabeled nodes. Fig. 6(a) shows the constructed network with the toy example and Fig. 6(b) shows the classification results by thresholding the predicted value (if $f > 0$ or not). As



(a) Constructed network of people (b) Classification results of power group

Fig. 6. Classification process of toy example.

a consequence, Persons E, G, and I now have predicted labels for their power groups, "+1", "+1", and "−1", respectively.

5. Results and discussion

In this section, we first explore the results of network construction with a specific time marker from the AGS and label setting using agendas from the AJD. We then present the classification performance using the proposed method and show comparisons with other algorithms. Lastly, we discuss the enrichment of the study of the classification results with actual opinions for representative agendas of the AJD.

5.1. Results for network construction

For the first step of network construction, we used Seo Geojung (1420–1488), who published the AGS, as the time marker. By applying the flexible time-window for year 1466—the year Seo Geojung became a national bureau official (YTP_{TM})—we derived the upper and lower bounds as 1488 and 1443, respectively. Thus, the period 1443–1488 was set as the time window. Using the proposed method, contemporary people were set as nodes of the network. Numerically, a network of 1589 nodes was established. Among them, 446 people had records for their admission year to a national bureau (YTP) in the period while the remaining 1142 people, who are the siblings of those people, were also added. The edges were defined by the similarity matrix using (6). Some trivial edges whose values were less than threshold $\varepsilon = 5 \times 10^{-4}$ were removed. Consequently, the number of resulting edges were 36,965 relations for 1589 people. Fig. 7 illustrates the (sub-)network and the insert shows its magnification, centered on Seo Geojung. In the figure, the numbers shown on the edges indicate the degree of kinship along with the weight values between the associated nodes. The thicker edges represent a closer degree of kinship and therefore higher similarities.

For the labels, we collected the opinions of people for 126 agendas in the AJD from 362,161 articles (simply speaking, each of them is a web page) within the time window. The collected data consists of opinions for 74 people out of 1589 people in the network, in various forms that are not limited to agree/disagree. To orient them toward the time marker, we preprocessed the data into a binary form representing pros/cons with respect to his opinion for each agenda. For each agenda, the number of people expressing their opinions varied from a minimum of one person (11 agendas) to a maximum of 23 people (two agendas). Fig. 8 shows a typical result for 15 selected agendas. The participants per agenda vary and accordingly, hence the bar heights are different in the figure. Blue dotted bars and white bars indicate the number of people who agree and disagree with the time marker, respectively.

To assign power groups to each individual, we first defined two power groups: a) the group consisting of members whose opinions are similar to the time marker and b) the group consisting of members whose opinions are different from the time marker. Hence, the phi coefficient was calculated. Out of 74 people, 65 people were assigned to power groups while nine people were left undecided, i.e., they had no common agenda for expressing an opinion with the time marker, therefore, for them, the phi coefficient was $\varphi = 0$. For the undecided people, we compared the φ values with respect to each individual who are already in either group and assigned them to the group with the majority of similarities. Through the process, we assigned 56 people to one group (group 1) and 18 people (group 2) to the other. Table 3 summarizes the data used for the network construction.

5.2. Results for network construction

5.2.1. Performance of the proposed method

To check the performance of the proposed method, we applied graph-based SSL algorithm to a constructed human network for predicting the labels of the nodes. For the experimental setting, we differentiated the number of labeled data to see if more labels provided higher prediction accuracies. Eight sets of labels ($74 \times$ [0.05, 0.10, ..., 0.40]) of different sizes were used for prediction, and each experiment was repeated 100 times. For the tradeoff parameter μ, we obtained the optimal value through cross-validation over the range $\mu \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, which was 100. The performance was measured using the area under the receiver operating characteristic curve (AUC) (Allouche, Tsoar, & Kadmon, 2006; Fawcett, 2006; Powers, 2011; Provost & Fawcett, 1997; Shin & Cho, 2006). The AUC ranges from 0 to 1, where an AUC value of 0.5 corresponds to random guessing and a higher value indicates better prediction accuracy.

Fig. 9 shows AUC values with respect to an increase in the size of the label set. The figure shows that there exists an increasing trend of the AUC with respect to increases in the size of the label set. This implies higher performance for a higher number of labeled data. The maximum AUC obtained with the graph-based SSL algorithm on the constructed network was 0.62. This quantitatively shows that blood relations of people are informative when inferring the power mechanisms of people.

5.2.2. Comparison with other classification algorithms

We compared the performance of the proposed method to other classification algorithms: ANN (Abraham, 2005; Bishop, 1995) and SVMs with polynomial and Gaussian kernels (Cristianini & Shawe-Taylor, 2000; Kotsiantis et al., 2006; Schölkopf & Smola, 2002; Shin & Cho, 2007).

For the experimental setting, we used 74 labeled data with 40% as the training set and 60% as the test set. For each algorithm,

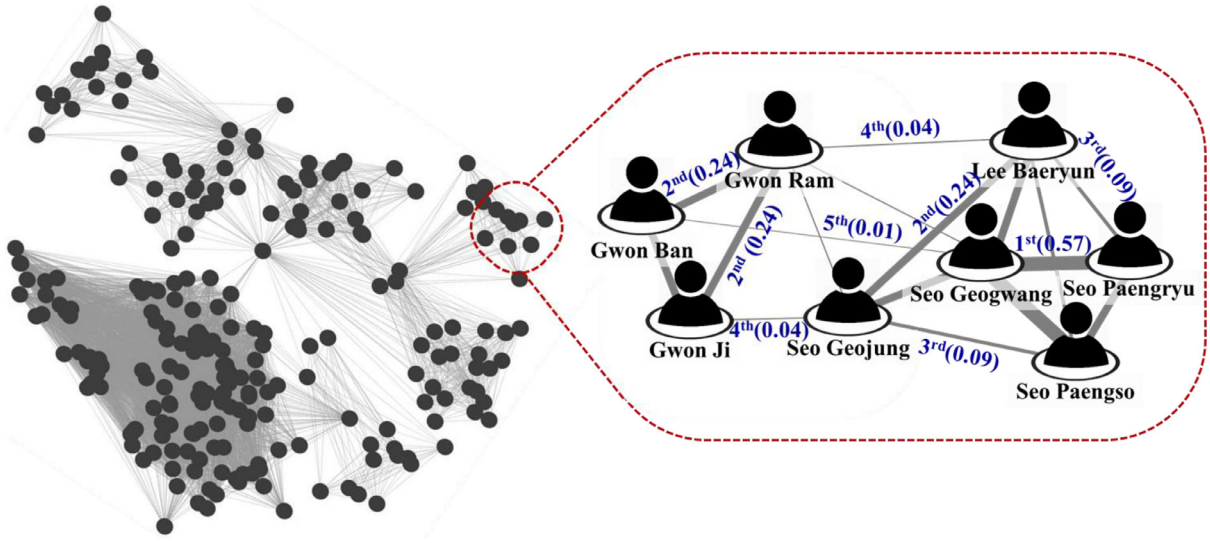


Fig. 7. Subset of the network for the Andong Gwon clan and magnification centered on the time marker Seo Geojung.

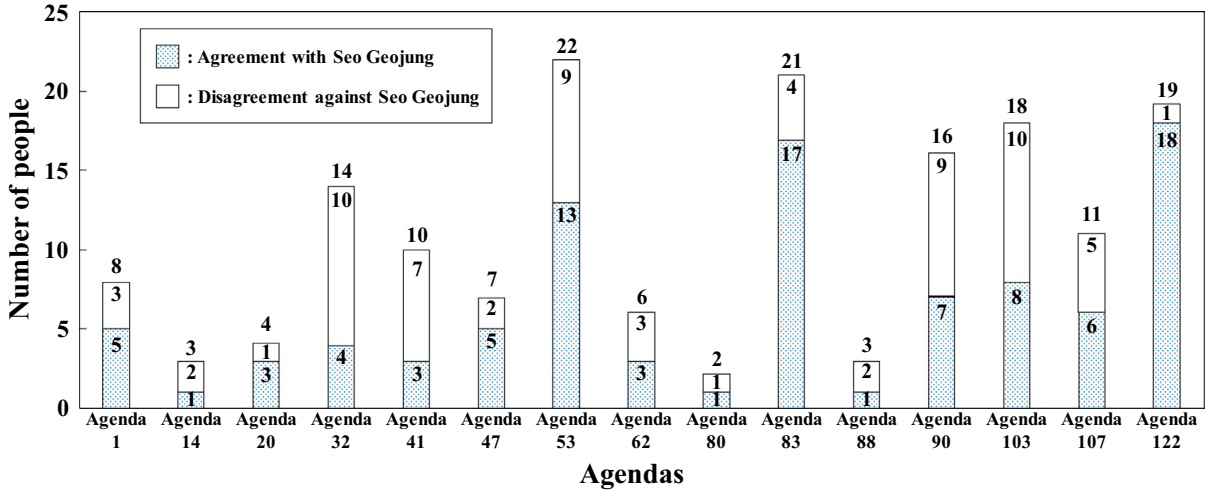


Fig. 8. Number of people showing positive or negative opinions on case-by-case agendas: the blue dotted bars and white bars represent the number of people who agree with and disagree with time marker Seo Geojung, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Summary of historical data used for the human network.

Data source	Total number of data	Number of data used for network construction		
		Nodes	Edges	Labels
AGS	10,243 people	1589 people	36,965 relations	74 (56:18)
AJD	362,161 articles	–	–	126 agendas

the experiment was repeated 100 times. Inputs were fed to both algorithms after encoding the genealogy tree into binary vectors. Each generation was represented as a 4-bit vector. For instance, to encode three generations of a subject, his grandfather becomes 0001-0000-0000, his father (the third son of grandfather) is 0001-0011-0000, and the subject himself is coded as 0001-0011-0001. Because the total number of generations included in the AGS is 21, the encoding produces 84-dimensional binary vectors. After removing the redundant dimensions (indicating "off-the period"), the resulting dimensions of the input became 44. For parameter setting, the tradeoff parameter μ was set as 100 for SSL as explained previously; the number of hidden nodes was set to 20 for ANN (44-20-2 MLP); the penalizing parameters of the SVMs were set to 10 and 1 for the polynomial SVM and Gaussian SVM, respectively. The poly-degree of the former was set to 1 whereas the kernel width was set to 0.0001 for the latter. These values were obtained through cross-validation. The performance was measured by the AUC and balanced correct-classification rate (BCR) (Shin & Cho, 2006). BCR treats the class imbalance problem by reflecting the balance between the different sizes of positive and negative classes. To calculate the BCR, we utilized the following equations:

$$S_e = \frac{Number\ of\ predicted\ positivess}{Total\ number\ of\ actual\ postives},$$

$$S_p = \frac{Number\ of\ predicted\ negatives}{Total\ number\ of\ actual\ negatives},$$

$$BCR = S_e \cdot S_p. \tag{8}$$

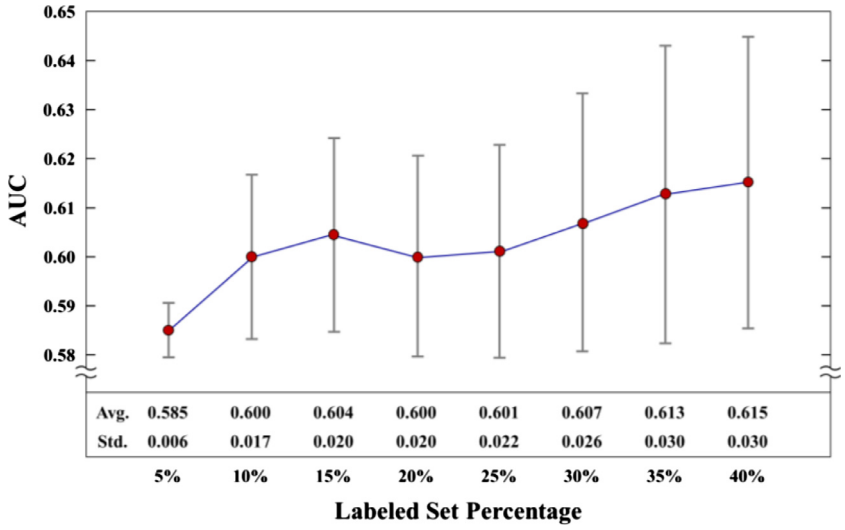


Fig. 9. Performance (AUC) of the graph-based SSL algorithm in power group classification.

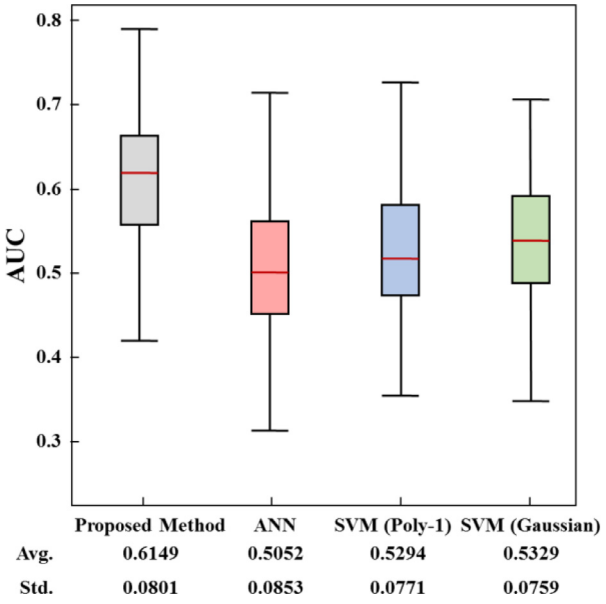


Fig. 10. Box-whisker-plot for performance variation across each algorithm: The red line inside the box indicates the median AUC for repeated experiments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

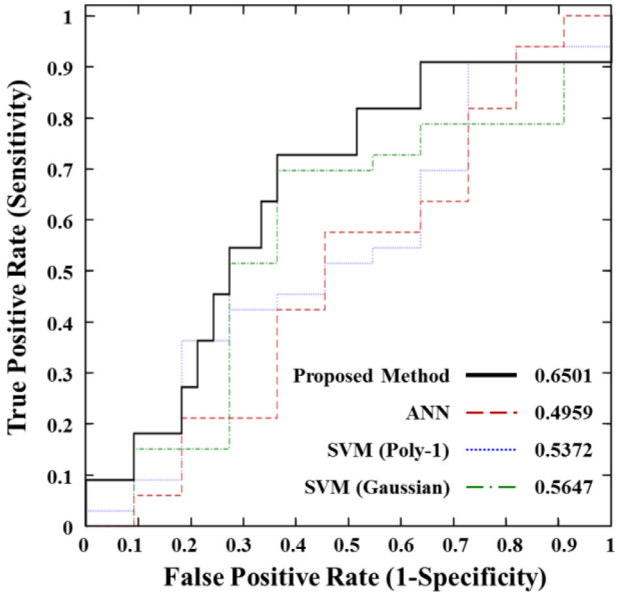


Fig. 11. Typical ROC curves for each algorithm: The closer the curve follows the left border and then the top border (i.e., gaining a larger area under the curve), the better performance of the classifier.

In (8), S_e and S_p are the sensitivity and specificity, respectively. Here, sensitivity refers to the true positive (group 1) rate and specificity refers to true negative (group 2) rate (Altman & Bland, 1994). The BCR ranges from 0 to 1, and higher value stands for better prediction performance. In our case, as the ratio of group 1 to group 2 was three to one (56:18), BCR was more suitable as a performance measure.

Fig. 10 presents a box-whisker-plot with the mean of the AUC and its standard deviation for each algorithm. The red line inside the box indicates the median AUC for repeated experiments. From the plot, we see that the proposed method outperforms other algorithms with highest median AUC, followed by SVM (Gaussian), SVM (Poly-1), and ANN. In addition, the results show that the proposed method achieves the highest mean AUC with 0.6149, followed by the same order as the median with AUC values of 0.5329, 0.5294, and 0.5052, respectively. It is worth noting that SVM (Gaussian), SVM (Poly-1), and ANN attained AUC values close to 0.5, which almost correspond to that of random guessing. This result conveys the fact that the proposed method incorporates the genealogy information into the task of power group classification relatively better than the other representative machine learning algorithms.

Fig. 11 shows a typical receiver operating characteristic (ROC) curve for each algorithm. For ROC curves, the closer the curve follows the left border and then top border (i.e., achieving larger area under the curve), the better performance of the classifier. Thus, this figure illustrates that the proposed method is more accurate than other possible models for power group classification.

Table 4 shows a confusion matrix for the experiment. From the table, we see that BCRs were 0.2066, 0, 0.1598, and 0 for SSL, ANN, SVM (Poly-1), and SVM (Gaussian), respectively. In addition, accuracies were 0.64, 0.75, 0.70, and 0.75, in the same order. From the results, we see that ANN and SVM (Gaussian) achieve the highest

Table 4
Typical confusion matrices of the four algorithms used for power group classification.

(a) Proposed Method

		Predicted Group 1	Predicted Group 2
Actual	Group 1	25	8
	Group 2	8	3

BCR = 0.2066

(b) ANN

		Predicted Group 1	Predicted Group 2
Actual	Group 1	33	0
	Group 2	11	0

BCR = 0

(c) SVM (Poly-1)

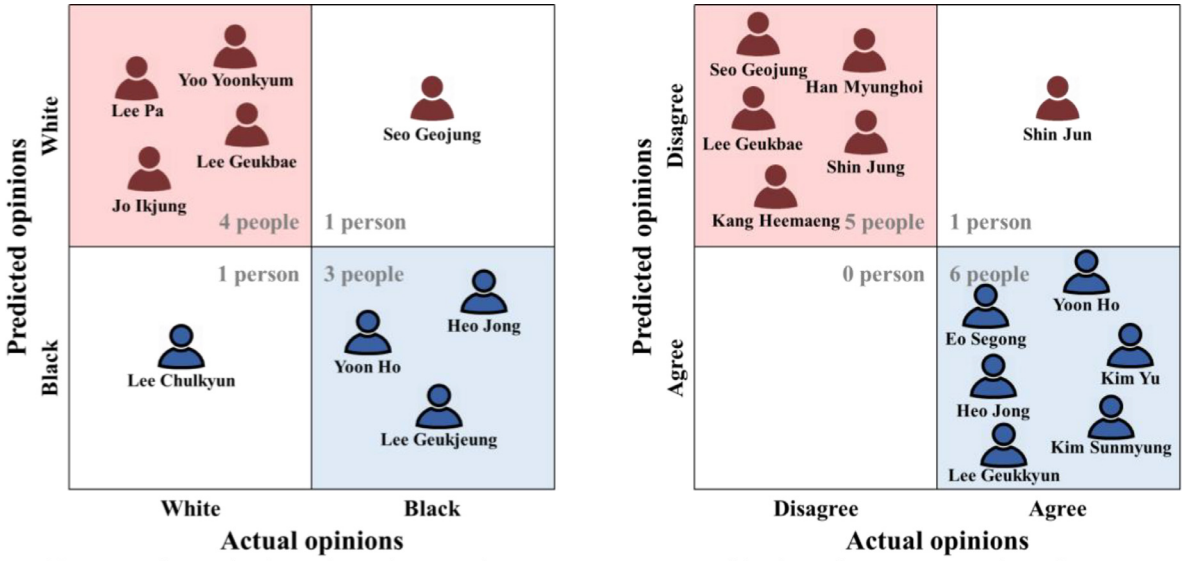
		Predicted Group 1	Predicted Group 2
Actual	Group 1	29	9
	Group 2	4	2

BCR = 0.1598

(d) SVM (Gaussian)

		Predicted Group 1	Predicted Group 2
Actual	Group 1	33	0
	Group 2	11	0

BCR = 0



(a) Appropriate attire for ambassador reception (b) Discussion on constructing a fort

Fig. 12. Results of the inference of opinions by classifying power group of people in Sunghwabo: In both cases, red people are classified as belonging to one power group and the blue people with thick borders are classified as belonging to the other. People with the same-colored background indicate those that have the same inferred and actual opinions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

accuracy. The measurement, however, does not take into account the class imbalance problem, as they classify all data points into one power group. Because we have more people for group 1, it is obvious that if a classifier predicts all data points as belonging to the associated group, then it achieves a high accuracy. By observing the BCR, which takes such issues into account, we see that the proposed method attains the highest value, which confirms the high performance of the proposed method.

5.2.3. Enrichment study

We further enriched the classification results with the actual opinions for representative agendas of the AJD. Eight agendas were chosen in which more than 10% of the people in the network (with labels) expressed their opinions. Twelve of these people were randomly selected and endowed with labels. The other 62 were used to measure the classification performance; therefore, their true labels were not given to the graph-based SSL algorithm and inferenced as a result.

Fig. 12 depicts the results of opinion inference for two specific agendas. As background information, Fig. 12(a) deals with the agenda of the appropriate attire to wear to an ambassador reception during the national funeral of Queen Junghee. In the Joseon Dynasty period, when there was a national funeral due to the death of a king or a queen, every official had to wear white colored attire. Soldiers, however, had brightly colored armor, so they had to choose from among black, white, and other colors. In the end, it was decided that the soldiers would wear black armor at the ambassador reception.

The agenda in Fig. 12(b) deals with installing a fort in Haenam (a region of Jeolla-do, Korea). A plan for installing a fort in the region was suggested to defend it more effectively from enemies. In terms of the installation and maintenance, however, opposing views were suggested because there were not sufficient human resources. After numerous discussions, the king decided not to install the fort. From these two example agendas, we can see that the number of people expressing their opinions is different for each case. These two examples seem to show that the proposed method produced a satisfactory inference on opinions.

From genealogy, Lee Geukbae, Lee Geukjeung, and Lee Geukkyun are siblings. But in the former agenda, Lee Geukbae and Lee Geukjeung were classified in different power groups with showing disparate opinions. Similarly, Lee Geukbae and Lee Geukkyun were split differently in the latter agenda. Since the proposed network is constructed from genealogy, splitting power group within a family can be seen anomalous. However, it is possible to see such cases since people have different opinions despite having close blood relations. These depicts such phenomenon. In conventional historical studies, siblings are considered to be in same power group. In contrast, the proposed method may shed light on discovering diverse perspectives.

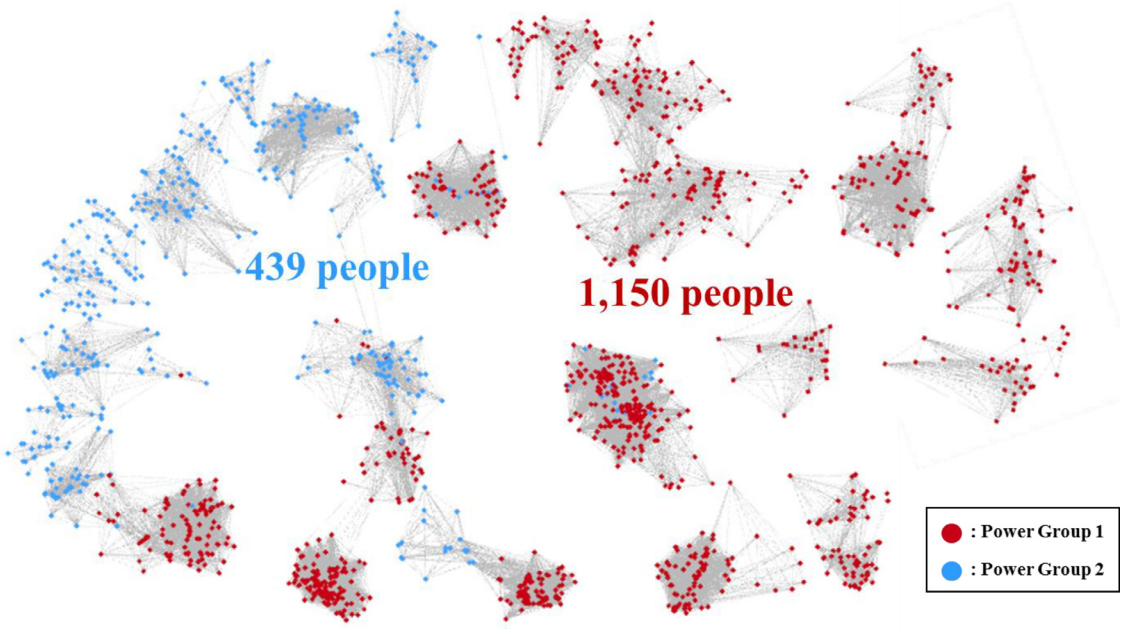


Fig. 13. Human network of the Andong Gwon clan (1589 people for the period 1443–1488).

After verifying the performance of the algorithm, we expanded the label inference to classify all of the 1589 people in the human network. In this setting, all of the 74 labeled people were set as labeled nodes. Consequently, 1150 people and 439 people were classified into each of the power groups. Fig. 13 shows the full human network of 1589 people. The colors of the nodes distinguish the two power groups.

This outcome is expected to rethink the topography of the political forces in an era according to the trend of decision-making by historical figures. In previous studies on power mechanism the historical figures have been classified not by rigorous analysis based on related historical documents but by binary conflict structure between two political forces like the Meritorious Elite and Neo Confucian in early Joseon Korea. In contrast, our new approach expands our insight regarding political forces into diversity, not uniformization like a priori study. In other words, this proposed method leads the historians to attempt to classify some historical figures who cannot confirm political stances into political forces different from the priori study. In fact, many historical figures classified into two power group in Fig. 13 had no clear information on their political opinion of each agenda. With the help of SSL, it is possible to classify most historical figures based on the links of the networks.

Moreover, the proposed method contributes to reduce time and effort by domain historians to probe and analyze historical big data to derive a holistic understanding of historical landscape, which features interactions among a diverse range of social groups. Considering new research circumstance deriving from historical big data, machine learning approach such as our method helps the historian to draw big picture of power mechanism in history over long term period of time.

6. Conclusions

In this study, power mechanisms were analyzed using a machine learning approach. Based on two historical documents, the AGS and AJD, a human network was constructed and the power mechanisms were inferenced by propagating power influences via blood-relation edges in the network. First, we chose people from the AGS to be included in the network by defining a period of contemporaries. A flexible time-window method was suggested for this task. The similarities between the people in the network were also proposed based on the degrees of kinship obtained from the tree structure of the genealogy. The labels that indicate the memberships of people to power groups were created by calculating values of the phi coefficient. For this task, we analyzed the opinions of people about major agendas in the AJD and defined friendly or conflicting relationships among them. By applying the graph-based SSL algorithm to the constructed network, which was partially labeled by the phi coefficient values, it was possible to quantitatively infer the power group of people included in the network. It amounts to around 1500 people in the period of 1443–1488 in the Joseon Dynasty. Using our proposed method, we achieved an AUC of 0.62. To take into account for the fact that we only used blood-relations to identify the power structure, the obtained results can be regarded as satisfactory. We believe that if we are given more diverse historical sources, such as materials on regionalism or on alumni, better performance can be achieved.

There are various algorithms for solving prediction and inference problems for large and various types of data. There are, however, limitations when applying the existing algorithms directly to the historical data. The significance of this paper lies in its insights about how to solve such difficult problem. The proposed methodology is an initiative work in history study and can be further applied to any of era in the genealogy. In addition, this study can serve as a new tool that not only analyzes the power relationships among historical figures but also provides synergy with existing qualitative methods. From the viewpoint of history, this study can bestow an insight upon overcoming the limitations of current studies that attempt to classify power groups solely based on historical facts. It can also provide a methodology to systematically understand the structure of power groups over long periods of time. Furthermore, we expect that this study will grant us an opportunity to study the structure of elite classes from pre-modern to modern eras in order to understand the overall structure of politics, society, and the economy. Not being unique to Korean history, politics, and society, the framework presented in this work can be similarly expandable to any era in any country.

There are, however, few limitations to our research. First, because we consider people in the genealogy as a whole in a

network-wise fashion, we do not reflect individual characteristics or historical aspects in our method. Second, the constructed network in this study only utilizes blood relations as the degree of kinship in the genealogy and does not reflect other various relations of people, such as regional or alumni relations. Furthermore, we limited our evaluation of the proposed method to a small portion of historical people and agenda. We believe that we could extend the scope of the research by obtaining more labeled data for power groups that would eventually lead to a higher performance of the proposed method.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (Nos. 2015S1A5B6037107 / 2015R1D1A1A01057178) and the Ajou University research fund.

References

Abraham, A. (2005). *Artificial neural networks. Handbook of measuring system design*.
Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology, 43*, 1223–1232.
Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal, 308*, 1552.
Bak, J., & Oh, A. (2015). Five centuries of Monarchy in Korea: Mining the Text of the annals of the Joseon dynasty. *LaTeCH, 2015*, 10.
Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university Press.
Boas, F. (1909). Determination of the coefficient of correlation. *Science, 29*, 823–824.
Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.
Chapelle, O., Scholkopf, B., & Zien, A. (2006). *Semi-supervised learning. 2006*. Cambridge, Massachusettes: The MIT Press.
Chung, J.-S., & Kim, H.-Y. (2011). Analysis of people networks in Goguryeo, Baekje, and silla dynasty silloks (in Korean). *The Journal of the Korea Contents Association, 11*, 474–480.
Chung, S. (1983). *A study on social movement in late Joseon dynasty (in Korean)*. Seoul: Iljogak Press.
Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
Dahl, R. A. (1957). The concept of power. *Behavioral science, 2*, 201–215.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*, 861–874.
Foucault, M. (1980). *Power/knowledge: selected interviews and other writings, 1972-1977*. New York: Pantheon.
Goldberg, A. B., & Zhu, X. (2006). Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the first workshop on graph based methods for natural language processing* (pp. 45–52). Association for Computational Linguistics.
He, J., Carbonell, J. G., & Liu, Y. (2007). Graph-Based Semi-Supervised Learning as a Generative Model. *IJCAI, 7*, 2492–2497.
Henry, R. B. (1935). *Genealogies of the families of the presidents*. Tuttle Company.
Kang, D. (1980). *History of korea invasion policy by japan (in* Korean*)*. Seoul: Hangilsa Press.
Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 119–127.
Kim, J., & Shin, H. (2013). Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *Journal of the American Medical Informatics Association, 20*, 613–618.
Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*, 159–190.
Lee, H.-s. (2004). Edward W. Wagner: The Father of Korean Studies in North America. *The Review of Korean Studies, 7*, 117–131.
Lee, J. Z., & Campbell, C. D. (2010). China multi-generational panel dataset, Liaoning (CMGPD-LN). *Data Sharing for Demographic Research (DSDR)*, 1749–1909.
Lee, K., Lee, J., Shin, H., Jung, M., & Yoo, Y. (1993). *Political form of* Korean *history (in* Korean*)*. Seoul: Iljogak Press.
Lee, S. (1979). *Formation of* Youngnam Sarim *faction (in* Korean*)*. Kyungsan: Youngnam University Press.
Lee, S. (2000). *History of political party strife in* Joseon *dynasty (in* Korean*)*. Seoul: Dongbang Media.
Lee, S. (2010). The impacts of birth order and social status on the genealogy register in thirteenth-to fifteenth-century Korea. *Journal of Family History, 35*, 115–127.
Lee, S. (2013). The impact of family background on bureaucratic reproduction in the thirteenth -to- fifteenth century Korea: A case study on the Kwon-ssi Sunghwabo (in Korean). *Daedong Munhwa Yeon'gu, 81*, 41–67.
Lee, S. (2016a). Conditions and potentials of Korean history research based on 'big data'analysis: The beginning of 'digital history'. *Korean Journal of Applied Statistics, 29*, 1007–1023.
Lee, S. (2016b). Towards a sustainable future for historical demography. In S. H. Koen Matthijs, & Jan Kok, Hideko Matsuo (Eds.), *The future of historical demography: upside down and inside out*. Acco Leuven.
Lee, S., & Son, B.-g. (2012). Long-term patterns of seasonality of mortality in korea from the seventeenth to the twentieth century. *Journal of Family History, 37*, 270–283.
Lee, T. (1985). *Political history revisit of* Joseon *dynasty (in* Korean*)*. Seoul: Taehaksa Press.
Lee, T. (1986). *A study on* Korean *social history (in* Korean*)*. Seoul: Jisik Sanup Publications.
Liu, Y., Dai, S., Wang, C., Zhou, Z., & Qu, H. (2017). GenealogyVis: A system for visual analysis of multidimensional genealogical data. *IEEE Transactions on Human-Machine Systems*.
Manabe, T. (1999). The digitized kobe collection, phase I: Historical Surface marine meteorological observationsin the archive of the Japan Meteorological Agency. *Bulletin of the American Meteorological Society, 80*, 2703–2715.
Manoff, M. (2010). Archive and database as metaphor: Theorizing the historical record. *Portal: Libraries and the Academy, 10*, 385–398.
Min, H.-g. (1974). Conservative powerful clans in late Goryeo dynatsty (in Korean). *Hanguksa 8.*. Seoul: National Institute of Korean History.
Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 195*, 1–405.
Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*, 37–63.
Provost, F. J., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. *KDD, 97*, 43–48.
Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing, 2016*, 67.
Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.
Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.
Sharma, S., Agrawal, J., Agarwal, S., & Sharma, S. (2013). Machine learning techniques for data mining: A survey. In *Computational Intelligence and computing research (ICCIC), 2013 IEEE international conference on* (pp. 1–6). IEEE.
Shin, H., & Cho, S. (2006). Response modeling with support vector machines. *Expert Systems with Applications, 30*, 746–760.
Shin, H., & Cho, S. (2007). Neighborhood property–based pattern selection for support vector machines. *Neural Computation, 19*, 816–855.
Shin, H., Hill, N. J., Lisewski, A. M., & Park, J.-S. (2010). Graph sharpening. *Expert Systems with Applications, 37*, 7870–7879.
Singh, Y., Bhatia, P. K., & Sangwan, O. (2007). A review of studies on machine learning techniques. *International Journal of Computer Science and Security, 1*, 70–84.
Song, J.-h., & Wagner, E. W. (2000). *Supplementation and annotation of* Mungwa *roster of* Joseon *dynasty (in Korean)*. Seoul: Dongbang Media CD-ROM.
Subramanya, A., & Talukdar, P. P. (2014). Graph-based semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 8*, 1–125.
Wang, J., Shen, X., & Pan, W. (2009). On efficient large margin semisupervised learning: Method and theory. *Journal of Machine Learning Research, 10*, 719–742.
Yi, W. (1991). *A study on* Korean *medieval society (in* Korean*)*. Seoul: Iljogak Press.
Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75*, 579–652.
Zhu, X. (2005). *Semi-supervised learning literature survey* (p. 1530). University of Wisconsin Madison Computer Sciences TR.
Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *ICML, 3*, 912–919.
Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3*, 1–130.