Identifying Guarantors of War Veterans Using Link Prediction Based on GraphSAGE: A Case of the Korean War

Abstract

Many veterans who participated in war now suffer from mental and physical health-related issues. In order to reward their sacrifice, each country provides veterans with various benefits and welfare. However, if their service records are not recognized due to the loss of related official documents, it is impossible to obtain these benefits. In such cases, a "buddy statement" can play an important role. However, many veterans have difficulty finding someone who can vouch for their participation in the war. To solve this problem, this study proposes a combined operations network in which veterans can find guarantors who participated in the same battle. If it is difficult to find a guarantor directly for reasons such as death, link prediction can be used to identify highly relevant guarantors in this network. We apply our proposed approach to Korean War data to train a combined operations network with GraphSAGE by sampling neighbors and using various kinds of aggregation functions. The comparison of the prediction performance to those of other baseline models shows the superiority of our proposed GraphSAGE based approach.

Introduction

Many soldiers who participated in war have sacrificed themselves for their country. Each country honors those soldiers by awarding them medals and providing them with various kinds of welfare and benefits (Casler and Fosmire 2019). However, there are cases where veterans' service records are not recognized for various reasons, such as the loss or disappearance of service records. In the 20th century, numerous wars broke out, including the Korean War and the Vietnam War. Unlike the present, when all records are digitally databased, in the past many records were written by hand and stored physically. Therefore, these records could easily be lost, either naturally or due to other wars. For this reason, if a service record is not found, each country acknowledges the participation in wars in the following manner.

In the United States, the National Personal Records Center (NPRC) handles personnel records for the U.S. military. Due to a fire at the NPRC on July 12, 1973, many materials were lost, including approximately 80% of Army records from 1912 to 1960 (Walter and Evans 1974). Some records have been restored, but there are still people for whom there is no data for their veteran status to be recognized. To this end, the U.S. Division of Veterans Affairs [1 allows for the recognition of war efforts by referring to the following documents:

- Statements from service medical personnel;
- Certified "buddy statements" or affidavits from fellow service members who witnessed the veteran's injury or illness;
- Military accident and policy reports;
- Examination reports related to employment or insurance;
- Letters or photographs from the veteran's time in the service;
- Prescription records; and
- Photocopies of any service treatment records or medical reports from any private hospitals, clinics, or doctors who treated the veteran during service or shortly after separation.

Since Korea's liberation from Japanese colonial rule in 1945, Korea has participated in two major wars—the Korean War and the Vietnam War. Especially in the case of the Korean War, many records were lost because data were stored physically. Therefore, the Ministry of National Defense and the Ministry of Patriots and Veterans Affairs recognize participation in the war by compiling the following data:

- Registration application for war veterans;
- Confirmation of participation in the war and certificate of removal from the family registry;

1 https://www.va.gov/records/get-military-service-records/reconstruct-records/

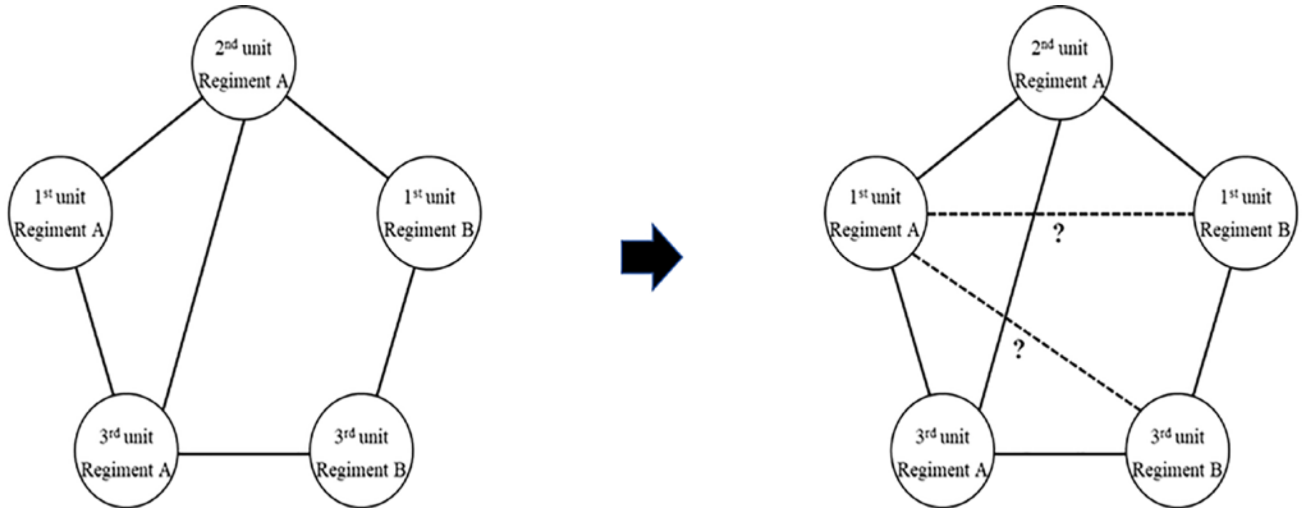

Figure 1: Identification of guarantor for war veterans using link prediction in a combined operation unit network

- Original documents indicating participation in the war or buddy statements; and
- Additional requirements for each veteran.

In each country, there are still people who are not recognized as veterans, even though various documents are collected to account for the absence of records. Most veterans have difficulty obtaining a guarantor's statement among the documents they submit. In the case of battles that occurred decades ago, there may not be many survivors, so it can be almost impossible to find comrades belonging to the same unit.

However, there is currently no research to solve the difficulties arising from trying to be recognized as a veteran. Many previous studies have dealt with veterans' issues such as social life and mental health after experiencing war, but none have paid attention to difficulties in the recognition process. Studies have identified the problem of finding veterans, but there is no acknowledgement of soldiers without records (Ra 2017). Therefore, herein, we take note of the difficulties of veterans arising from the absence of records, which are not covered in existing research, and present potential solutions.

The goal of this paper is to propose a database framework to recommend suitable guarantors for veterans, as illustrated in Figure 1. The framework addresses the case wherein it is difficult to find a guarantor directly for reasons such as death. This study proposes a link prediction model to identify potentially relevant guarantors using a combined operation unit network trained with GraphSAGE, which involves the process of sampling and aggregating neighbors. The GraphSAGE model has the advantage of being applicable to other similar war networks without additional training.

Related Work

Veterans Affairs

There have been several efforts to solve difficulties often encountered by veterans. However, most studies focus on welfare benefits, such as resolving social problems after retirement and treating post-traumatic stress disorder (PTSD) caused by war. Kang et al. (2003) pointed out that the ratio of PTSD and chronic fatigue syndrome (CFS) in U.S. soldiers as a result of the Gulf War was relatively high, and emphasized the need for testing and treating veterans for PTSD and CFS. Williamson et al. (2018) focused on the need for further research on mental health and treatment of veterans aged 65 and older.

Casler and Fosmire (2019) addressed the difficulties experienced by veterans with respect to their communities, families, and mental health, and stressed the need for improvement in medical services by analyzing the needs of veterans. Nelson et al. (2015) focused on the difficulties veterans experience in employment due to physical and mental problems, and emphasized the necessity of research taking into account the employment-related demands of veterans to resolve the problem.

There was also a study that emphasized the difficulties experienced by veterans who were not recognized as veterans due to lack of records. Nam (2013) focused on the difficulties experienced by non-regular soldiers who participated in the Korean War. This study emphasized the problems of non-regular soldiers who were not recognized as veterans due to insufficient evidence. Jeong and Kim (2018) also observed the difficulties experienced by non-regular soldiers without sufficient records. Recommendations were made for the compensation and treatment of non-regular soldiers who had to sacrifice a lot, such as being deprived of educational opportunities due to their participation in the war as minors. However, neither study could provide an adequate solution for veterans who had no records of their participation in the war. As represented in many studies, veterans need mental, physical, and economic welfare. However, many veterans still struggle because they are not recognized for their participation in the war due to the absence of records.

Graph Neural Network

Scarselli et al. (2009) proposed a graph neural network (GNN) to express the underlying relationships between nodes. Since then, many GNN methods have been developed and applied to problems in various domains. Kipf and Welling (2016) contrived a graph convolutional network that generalized the concept of applying a convolution filter to graph-structured data. This method has been applied to many problems such as text classification (Yao, Mao, and Luo, 2019) and forecasting the demand of a bike sharing service at the station level (Kim, Lee, and Sohn, 2019). Fout et al. (2017) predicted the interface between proteins by applying GCN in drug design and invention. However, previous studies applied a transductive framework that is not suitable for analyzing inductive problems. To compensate for this shortcoming, Hamilton, Ying, and Leskovec (2017) proposed an inductive framework called GraphSAGE. Generating node embeddings, instead of training all neighborhood nodes, GraphSAGE conducts learning through sampling in the local neighborhoods and aggregating their features.

Furthermore, GraphSAGE trains the node-invariant weights to aggregate the sampled neighbors and a source node. Moreover, sampling neighbor nodes helps to learn robust representations from changes in neighbor structures or on entering new nodes. Such node-invariant weights and sampling procedures give GraphSAGE the ability to learn inductive representations of the nodes, making it an effective method to extend to embedding unseen nodes or unseen graphs.

Data and Methodology

This study proposes a solution to veterans who are suffering from a lack of records following the framework displayed in Figure 2.

In the absence of original documents about participation in the war, a buddy statement is required to identify participation. Thus, the participant's statement about the battle or operation must match those of the guarantor(s) for them to be recognized as a war veteran.

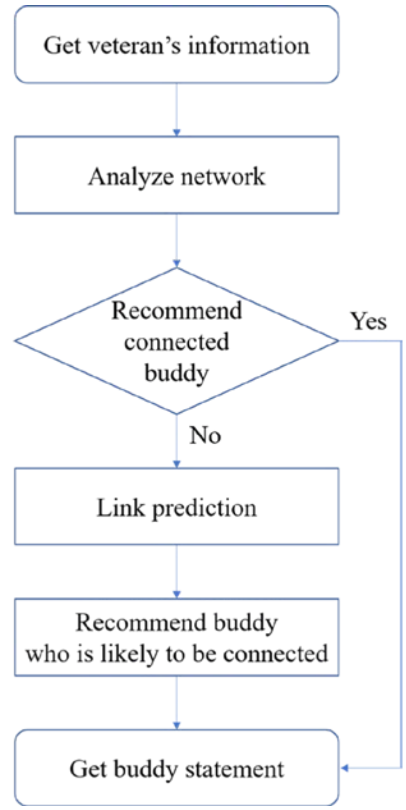

Figure 2: Framework flowchart

Previously, the process of receiving a buddy statement to prove participation was completed by the veterans themselves. In fact, there is no way to find a guarantor for a buddy statement unless they belong to the same unit as the veteran. Unlike previous frameworks, our framework can recommend a person who can provide a buddy statement, even if there is no direct connection to the veteran (neither belonging to the same unit nor participating in the same operation), by predicting the possibility of potential connection.

Data

In this paper, we propose a model that finds highly relevant guarantors through link prediction, taking the Korean War as an example. The Korean War was fought for three years, one month and two days, from June 25, 1950 to July 27, 1953. During this entire period, a total of 376 major battles and combined operations between forces were officially executed (Cho et al. 2017). After the Battle of Inchon on September 15, 1950, the United Nations Forces pursuit operation was carried out, resulting in frequent combined operations. In November 1950, due to the unexpected intervention of the People's Volunteer Army of China, retrograde operations were implemented, and combined operations became rare (Cho et al. 2017). Therefore, in this study, only combined operation data from September 16, 1950 to October 31, 1950 were analyzed.

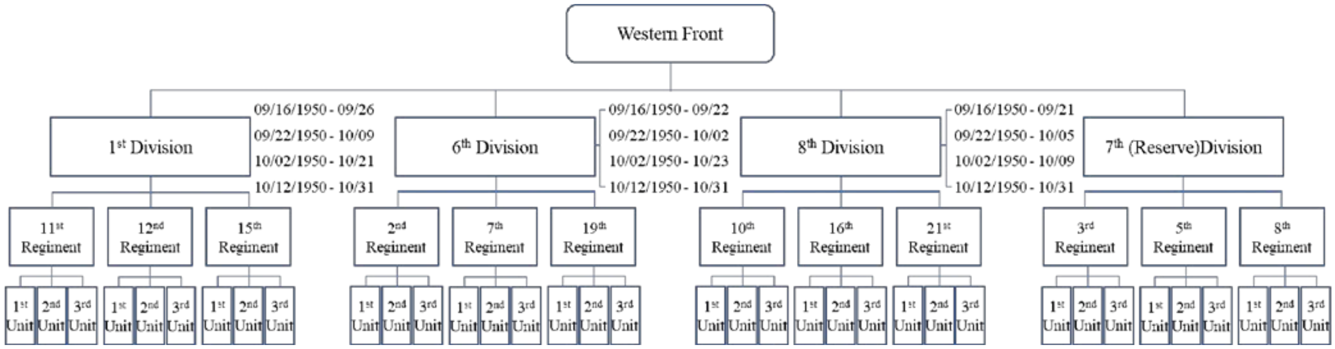

Figure 4: Western Front of ROKA (Republic of Korea Army), 1950

Figure 3: Eastern Front of ROKA (Republic of Korea Army), 1950

The details of the combined operation data were collected through Min (2009) to construct a network for recommending guarantors. By analyzing operational orders and movement paths between Army units included in the data, we identified cases in which combined operations between units such as shift changes and battles were executed. In particular, the movement route, base occupation, and advancement of each unit were converted into daily data. The combined operation data of the Western front (1st, 6th, 7th, and 8th divisions) and the Eastern front (capital and 3rd divisions) were extracted (see Figures 3 and 4). There were a total of 90 combined operations between the units during the defined period of study. Some of the combined operations were conducted between different regiments. For example, the recapture operation of Yongbyon was conducted between the 3rd unit of the 19th regiment in the 6th division, and all units of the 15th regiment in the 1st division, on October 24, 1950.

Soldiers in the war did not necessarily participate in all battles of their units. Therefore, based on the unit information of the soldiers killed, the number killed in each division was identified. Next, the period from September 16, 1950 to October 31, 1950 was classified into four smaller periods based on the dates when the number of soldiers killed per division per day changed significantly. For instance, the full period over which the Capital Division operated was split at September 21, October 1, and October 11, as displayed in Figure 4. The 7th Division was organized as a reserve division during this period, and due to infrequent combined operations, the period was not split for this division.

Therefore, the capital, 1st, 3rd, 6th, and 8th divisions were each classified into 15 ($= 24 - 1$) nodes depending on time periods. In addition, each division that participated in the operation during the period consisted of three regiments, and each regiment consisted of three units. Therefore, 675 (5 divisions×3 regiments×3 units×15 periods) nodes were created to include all divisions except the 7th, for which 9 (3 regiments×3 units) nodes were created.

Accordingly, a bipartite matrix was derived with two sets of nodes—684 unit nodes and 90 operation nodes. There are a total of 3342 edges between the unit nodes and operation nodes. We consider an undirected network that represents the relationship between unit nodes and operation nodes, depending on whether the units participated in the operations. The result of visualizing the bipartite matrix using Gephi, a network analysis and visualization package, is shown in Figure 5.

In addition, we consider the casualty information for nodes in this network. A total of 588,000 casualties of the Republic of Korea Armed Forces and 538,000 casualties of the UN Forces occurred during the Korean War. Therefore, we reflect the number of deaths in our model.

We collected data related to 191,451 Korean soldiers killed from the information retrieval system in the War Memorial of South Korea. The data consists of information on

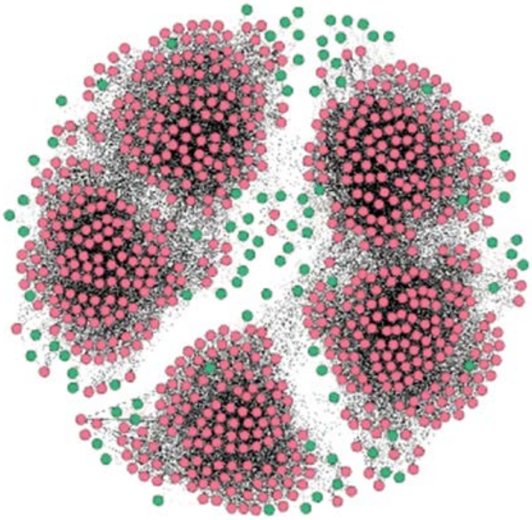

Figure 5: Visualization of Bipartite Network

the names, military classifications, units, ranks, service numbers, dates of birth, places of birth, dates of death, places of death, and relations of the soldiers who were killed. We use the daily casualties of each regiment for the node feature of Army unit information, as the casualty information was not available at the unit level.

For link prediction between two nodes, the adjacency matrix was derived from its unit node projection. We consider the number of operations in which two units participated together as the edge weight. The adjacency matrix representing the combined operation unit network of the Korean War is shown in Table 1. The number of edges between unit nodes is 32967, and the average weight is 2.3526. The average degree is 98.2, and the minimum is 15. The unit nodes contain unit information such as division, regiment, and the number of deaths as features.

	1	2	3		683	684
1	0	8	5		0	0
2	8	0	5		0	0
3	5	5	0		0	0
				⋱		
683	0	0	0		0	1
684	0	0	0		1	0

Table 1: Adjacency Matrix

GraphSAGE

In this study, the graph is represented as $G = (V, E)$, where V is a set of nodes with $|V| = n$ and E is the set of edges. Each node represents a unit and the weight of the edge A_{ij} between two nodes i and j, gives m, which is the number of battles and operations within which the units participated together.

` Let x_v $(\forall v \in V)$ be node features. In our case, each node has 29 node features, including unit information. Assuming that there is a parameter of K depth, which is used for aggregating information from other nodes, we set a weighted matrix W_k, $\forall k \in \{1, \cdots, K\}$, which is used for propagating information to the next layer. Let h_v^k be the embedding of node v after the kth layer, with $\mathrm{h}_v^0 = x_v$, the first aggregation step depending on node features. For $k = 1, \cdots, K$, repeat aggregations and concatenations are performed as follows to perform embedding.

$$\mathrm{h}_{N(v)}^k = AGGREGATE_k(\{\, \mathrm{h}_v^{k-1}, \forall v \in N(v)\})$$
$$\mathrm{h}_v^k = \sigma\left(W_k \cdot CONCAT\left(\mathrm{h}_{v-1}^k, \mathrm{h}_{N(v)}^k\right)\right)$$

where $AGGREGATE$ indicates an aggregating function that can be the mean, long short-term memory (LSTM), or pooling function, for example. Using these functions aggregates the neighbors of node v (denoted $N(v), \forall v \in V$). GraphSAGE concatenates two vectors $\mathrm{h}_{v-1}^k, \mathrm{h}_{N(v)}^k$ and multiplies them with W_k, reducing the dimension. Finally, the node embedding vector goes through a nonlinear activation function σ. We use RELU as the activation function.

Experiments

We compare the proposed GraphSAGE with the GCN along with the following basic methodologies, where $w(x, y)$ is the weight between nodes x and y, and $\Gamma(x)$ represents neighbor nodes of x:

- **Adamic–Adar (AA)** calculates the similarity of two nodes as (Adamic and Adar 2003)

$$\boldsymbol{WAA(x,y)} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) \cdot w(y,z)}{\log\left(\sum_{c \in \Gamma(z)} w(z,c)^2\right)};$$

- **Resource allocation (RA)** reflects the sum of inverse degrees of the nodes in the common neighbors (Tao, Linyuan, and Yi-Cheng 2009),

$$\boldsymbol{WRA(x,y)} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) \cdot w(y,z)}{\sum_{c \in \Gamma(z)} w(z,c)^2};$$

- **Common neighbors (CN)** represents the number of neighbors that two nodes have in common,

$$\boldsymbol{WCN(x,y)} = k^{1-\alpha} \cdot s^{\alpha}$$
where $k =$ Number of $\Gamma(x) \cap \Gamma(y)$ and
$$s = \sum_{z \in \Gamma(x) \cap \Gamma(y)} w(x,z) \cdot w(y,z);$$

- **Jaccard Coefficient (JC)** compares elements in the union with those in the intersection (Jaccard 1901).

$$\boldsymbol{WJC(x,y)} = \frac{s}{r-s} \ \ where$$
$$r = \sum_{a \in \Gamma(z)} w(x,a)^2 + \sum_{b \in \Gamma(z)} w(y,b)^2.$$

Model	Precision	Recall	F1 score
AA	0.8775 (±0.0115)	0.7824 (±0.0066)	0.8271 (±0.0054)
RA	0.9384 (±0.0076)	0.8062 (±0.0069)	0.8673 (±0.0051)
CN	0.8399 (±0.0178)	0.7813 (±0.0061)	0.8094 (±0.0060)
JC	0.9178 (±0.0139)	0.7667 (±0.0072)	0.8354 (±0.0072)
GCN	0.7389 (±0.0738)	0.9212 (±0.0514)	0.8160 (±0.0398)
SAGE-Mean	0.8884 (±0.0397)	0.9624 (±0.0159)	0.9231 (±0.0171)
SAGE-Mean pooling	0.8747 (±0.0401)	0.9628 (±0.0197)	0.9171 (±0.0154)
SAGE-Max pooling	0.7221 (±0.0606)	0.9660 (±0.0231)	0.8257 (±0.0379)

Table 2: Performance comparison to other baseline methods

Experimental Setup

In our experiment, we implement our model in Tensorflow. To evaluate the link prediction task in our war dataset, we separate the dataset into training and validation datasets. Moreover, the same number of positive and negative node pairs are extracted. In detail, we separate the edges of the war dataset by 90:10. Ten percent of the edges are regarded as positive pairs in the validation dataset, while the rest of the edges are considered those in the training dataset. Negative pairs are randomly sampled, and the number of negative pairs is the same as the positive pairs for each training and validation dataset. For baselines, we divide the entire edges into the training set and validation set by 90:10. The GraphSAGE-based models are conducted with the Adam optimizer (Kingma and Ba 2014). We use binary cross-entropy loss for supervised learning.

We also compare three variants of aggregator functions for the GraphSAGE as mean aggregator, mean-pool aggregator, and max-pool aggregator, except the LSTM aggregator because LSTM architecture processes inputs in sequence.

There are various hyperparameters that need to be set to use GraphSAGE. First, we set the depth K=2 as recommended by Hamilton, Ying, and Leskovec (2017) with a training batch size of 256. We perform parameter sweeps over learning rates {0.01, 0.001, 0.0001} as the step size for minimizing loss; 1-hop and 2-hop sample sizes are both {5, 10, 15}. As every node in our war dataset has at least 15 neighbors, we set the sample sizes as less than or equal to 15. Moreover, two cases of dropout probability are set as {0.3, 0.5}. The optimal hyperparameters are obtained by a grid search. We conduct each experiment 10 times with random seeds and compare the average performance. The best performance is obtained when learning rate = 0.01, 1-hop sample size = 15, 2-hop sample size = 15, and dropout = 0.3. For GCN, we use a softmax function for activation and train 300 epochs using Adam optimizer with the learning rate of 0.01 as Kipf and Welling (2016) recommended.

Results

Table 2 shows the results for link prediction, which consist of the precision, recall, and f1 scores of all models. As can be seen, the performance of the GraphSAGE method is better than the baseline methodologies. RA shows the best results for all metrics out of the baseline methodologies, however all these methodologies show similar results. As most of combined operations were carried out within the same regiment, our network is locally compact and RA appears to outperform the other baselines. In the GraphSAGE models, the performance of both mean and mean-pooling models exceed 87%. The performance of max-pooling is poorer than any other model tested in this study.

Mean-pooling is more effective than max-pooling in graph representation learning (Xu et al., 2018). Max-pooling applies only the max value in aggregation. However, in our data, many nodes are similar because they often participate in the same operations, so max-pooling was found to be not as effective as expected in our model.

In addition, GraphSAGE shows better performance than GCN. Unlike GraphSAGE, GCN tends to overfit due to learning all nodes.

Implementation

In sum, guarantors can be identified in the following order utilizing the GraphSAGE based link prediction:

- Recommend guarantors in the same Army unit (A);.
- If no one is in the same Army unit, recommend guarantors who participated in the same combined operations (B, E);
- If no one participated in the same combined operations, recommend highly relevant guarantors using the link prediction based on GraphSAGE (C, D in sequence, assuming that the probability of link prediction with C is higher than that with D) as illustrated in Figure 6.

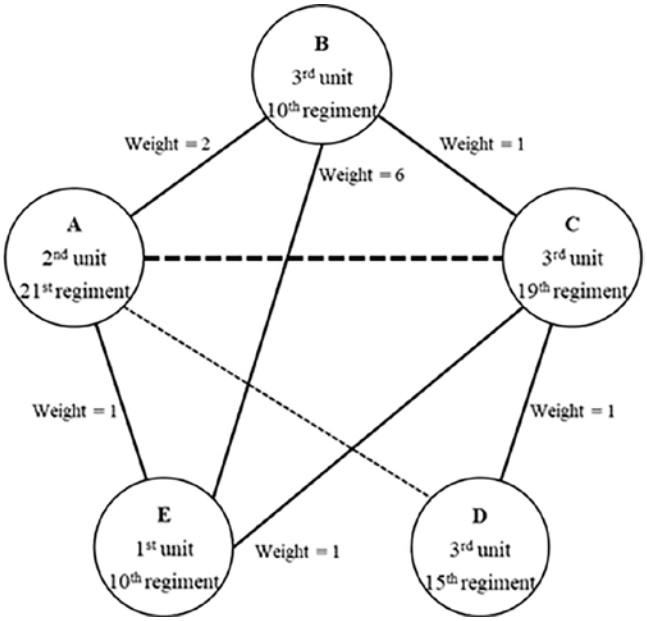

Figure 6: Identification of guarantor in the network

Conclusion and Discussion

The focus of this study was to resolve the difficulties experienced by veterans in the process of obtaining a buddy statement in the absence or loss of their war records. Our proposed combined operation unit network enables the recommendations of veterans who participated in the same operations or belonged to the same units. In addition, link prediction can be used to find a solution when no one can be found from a directly related link. By embedding nodes using GraphSAGE, we predicted the probability of link connection. The experimental results verified the accuracy and effectiveness of the proposed framework. As the inductive methodology is used, it has the advantage of being further utilized without additional learning when new operation nodes are added to the network.

Here in, a network was constructed using operation and battle data in the Korean War, and a unit-level network was formed. When the Personal Information Protection Act is relaxed in the future, further detailed network research may be conducted using the information of individual veterans.

In addition to GraphSAGE, Graph Attention Network proposed in Veličković et al. (2018) can be applied to our case using self-attention mechanism for setting the weight between nodes. These topics are left for further research areas.

References

Adamic, L. A., and Adar, E. 2003. Friends and neighbors on the Web. *Social Networks* 25(3): 211-230

Casler, M.; Fosmire, A.; and Klein, D. 2019. Support for Veterans: Community Reintegration, *Family and Mental Health Needs (Master's Thesis)*. Utica College, ProQuest Dissertations & Theses Global 59.

Cho, S. H.; Son, K. S.; Park, J. S.; Lee, S. H.; Bae, A. S.; Lee, G. S.; and Kim, H. G. 2017. The Korean War: Major Battles. Institute for Military History, Seoul, Korea

Fout, A.; Byrd J.; Shariat, B.; and Ben-Hur, A. 2017. Protein Interface Prediction using Graph Convolutional Network. *Proceedings of the 31st International Conference on Neural Information Processing Systems* 30: 6533-6542.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. *Neural Information Processing Systems* 2017: 1024–1034.

Jaccard, P. 1901. Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37(142): 547-579

Jeong, M. K., and Kim, S. Y. 2018. Korean War Veterans Experience of War and Meaning in Life. *Critical Social Welfare Academy* (58): 243-278.

Kang, H. K.; Natelson, B. H.; Clare M.; Mahan, C. M.; Lee, M. Y.; and Murphy, F.M. 2003. Post-Traumatic Stress Disorder and Chronic Fatigue Syndrome-like Illness among Gulf War Veterans: A Population-based Survey of 30,000 Veterans. *American Journal of Epidemiology* 157(2): 141-148.

Kim, T. S.; Lee, W. K.; and Sohn, S. Y. 2019. Graph convolutional network approach applied to predict hourly bike-sharing demands considering spatial, temporal, and global effects. *PloS one* 14(9). e0220782.

Kingma, D. P., and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv*:1412.6980.

Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations* (ICLR).

Min, H. K. 2009. Korean War History: Battle of Inchon and Counterattack. Institute for Military History, Seoul, Korea

Nam, K. K. 2013. The Necessity and It's Meaning of Compensation for Irregulars Contributor in the Korean War. *Korean Association of Unification Strategy Unification Strategy* 13(3): 9-32.

Nelson, C. B.; Abraham, K. M.; Miller, E. M.; Kees, M. R.; Walters, H. M.; Valenstein, M.; and Zivin, K. 2015. Veteran Mental Health and Employment: The Nexus and Beyond. *War and Family Life*: 239-260.

Park, D. C. 2014. The Korean War in statistics. Institute for Military History, Seoul, Korea

Ra, M. K. 2017. Korean Veterans' Policy through the Case of the Korean War Veterans. *Journal of Patriots and Veterans Affairs in the Republic of Korea* 16(4): 37-62.

Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; and Monfardii, G. 2009. The Graph Neural Network Model. *IEEE Trasactions on Neural Networks* 20(1): 61–80.

Tao, Z.; Linyuan, L.; and Yi-Cheng, Z. 2009. Predicting missing links via local information. *The European Physical Journa*l B 71: 623–630

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations* (ICLR)

Walter, W. S., and Evans, W. 1974. The National Personnel Records Center (NPRC) Fire: A Study in Disaster. *THE AMERICAN ARCHIVIST* 37(4): 521-549.

Williamson, V.; Stevelink, S.; Greenberg, K.; and Greenberg, N. 2018. Prevalence of Mental Health Disorders in Elderly U.S. Military Veterans: A Meta-Analysis and Systematic Review. *The American Journal of Geriatric Psychiatry* 26(5): 534-545.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks. In *International Conference on Learning Representations* (ICLR)

Yao, L.; Mao, C.; and Luo, Y. 2018. Graph convolutional networks for text classification. In *Association for the Advancement of Artificial Intelligence* (AAAI).