Extracting a Social Network among Entities by Web mining

YingZi Jin[1], Yutaka Matsuo[2], and Mitsuru Ishizuka[1]

[1] University of Tokyo, Hongo 7–3–1, Tokyo 113-8656, Japan
eiko-kin@mi.ci.i.u-tokyo.ac.jp, ishizuka@i.u-tokyo.ac.jp
[2] National Institute of Advanced Industrial Science and Technology
y.matsuo@aist.go.jp

Abstract. Social networks play an important role in the Semantic Web. Several methods exist to extract social networks among people such as FOAF aggregation, email analysis, and Web mining. In this paper, we expand the existing techniques for social network mining from the Web and apply them to obtain a social network for different entities. Especially, two types of networks are investigated in this study: firms and artists. Two technical improvements are made: relation identification and threshold tuning. Several evaluations emphasize the effectiveness of these methods. A social network of artists of the International Triennale of Contemporary Art (Yokohama Triennale 2005) was portrayed on the web site to facilitate navigation of the sources of artists' information. Our approach contributes to existing Semantic Web studies by cultivating the applicability of social networks from various domains.

1 Introduction

Social networks explicitly exhibit relationships (called *ties* in social science) among individuals and groups (called *actors*). They have been studied originally in social science since the 1930s. To date, a vast number of studies of social network analysis have been conducted. In the context of the Semantic Web, social networks are useful for trust calculation [7, 11], information sharing and recommendation [6, 12, 16], ontology construction [15], relations and relevance detection, e.g. COI detection [1], and so on.

Recent technologies have enabled us to obtain social network data: from e-mail archives, schedule data, and Web citation information. More recently, some sites such as LiveJournal and Tribe.net[3] generate numerous Friend-of-a-Friend (FOAF)[4] instances. Several studies have been launched to aggregate FOAF documents and to analyze the obtained social network which are connected by knows relationships [5, 15].

Another stream of research to obtain social network is to use a search engine: Simply put, we query a search engine about two names, then see how the two people are related. Co-occurrence of names on the Web is commonly used as proof of relation strength. Most of the studies in this way to extract social networks while targeting researchers or students [9, 13, 14]. This study is intended

[3] www.livejournal.com, and www.tribe.net.
[4] www.foaf-project.org

to expand current social network mining from the Web in order to apply it to other groups of entities than researchers. Specifically in this paper, we target two types of entities: artists of contemporary art and famous firms in Japan. Technically speaking, the expansion is not straightforward: Artists and researchers have different characteristics. Consequently, the appropriate Web mining methods are different. For the firm network, more improved algorithms for relationship identification are necessary. The artist network requires a sophisticated network generation from relational data using threshold tuning.

The contribution of this paper to this field is summarized as follows:

- We expand social network mining from the Web so that is applicable to various domains. Two major improvements are proposed and described; relation identification and threshold tuning.
- We show examples and evaluations for firms' and artists' networks. Our system was operated on the Web site for the International Triennale for Contemporary Arts (Yokohama Triennale 2005), to navigate users using the extracted social network of artists. We briefly overview the site.

This paper is organized as follows. The following section describes related studies and motivations. Section 3 investigates different appearance of entities on the Web, addresses our ideas to obtain various social networks from the Web. Sections 4 and 5 introduce our case study, which specifically addresses a complex and inhomogeneous community, and which uses firms and artists as examples. Application and discussion are shown in section 6 before we conlude the paper.

2 Related Works

In the mid-1990s, H. Kautz and B. Selman developed a social network extraction system called the *Referral Web* [9]. The system uses a general search engine to retrieve Web documents that include a given personal name. The names of other individuals are extracted from documents. In Referral Web, the strength of relevance of two persons, X and Y, is estimated by putting a query X AND Y to a search engine: If X and Y share a strong relation, we can usually find much evidence on the Web such as links found on home pages, lists of co-authors in technical papers, organization charts, and so on.

Recently, P. Mika developed *Flink*, a system for extraction, aggregation and visualization of online social networks for the Semantic Web community [14]. A social network of 608 researchers from both academia and industry is extracted and analyzed. The Web mining component of Flink, similarly to that in Kautz's work, employs co-occurrence analysis. The strength of the association between individuals was calculated by Flink using the Jaccard coefficient $n_{X\cap Y}/n_{X\cup Y}$, where $n_{X\cap Y}$ represents the number of hits by query X AND Y and $n_{X\cup Y}$ represents the number of hits by query X OR Y. The two individuals will be considered to have some relation if the value is greater than a certain threshold.

POLYPHONET also uses a search engine to measure the co-occurrence of names [13]. Actually, several co-occurrence measures have been compared, including the matching coefficient, mutual information, Dice coefficient, Jaccard

coefficient, overlap coefficient (aka Simpson coefficient), and cosine. The overlap coefficient $n_{X\cap Y}/\min(n_X, n_Y)$ performs best according to the experiments in the study. POLYPHONET was operated at several AI conferences in Japan[5] and at an international conference[6] to promote participants' communication.

A. McCallum and his group [2, 4] present an end-to-end system that extracts a user's social network. That system identifies unique people in e-mail messages, finds their homepages, and fills fields of the contact address book as well as the other persons' name. Links are placed in the social network between the owner of the web page and persons discovered on that page. A newer version of the system targets at co-occurrence information on the entire Web, integrated with name disambiguation probability models.

Name disambiguation is an important problem for social network mining. To date, several studies have produced attempts at personal name disambiguation on the Web [2, 3, 10].

3 Extraction Methods for Different Domain Entities

3.1 Problem of Existing Methods

The previous section presented an overview of social network mining studies. Co-occurrence-based approaches using a search engine are effective for the researcher domain. A question demands resolution: Are algorithms applicable to different kinds of entities? We would answer no for a couple of reasons.

For the first reason, co-occurrence-based methods become ineffective when two target entities co-occur universally on many Web pages. For example, when we want to find out two firms' relations from the Web and submit a query *Matsushita* AND *JustSystem* [7] to a search engine, we get as many as 425,000 pages, for which the Jaccard coefficient is 0.031, but this figure seems unreliable considering the probable predominance of coincidental page inclusions. Actually, this method does not work well for firms: it provides an error-ridden network.

For the second reason, co-occurrence-based methods work ineffectively when extracting a social network within different (or inhomogeneous) communities. For example, two Japanese artists *Taisuke Abe* and *Jun Oenoki* have no former relations, but the overlap/Jaccard coefficient is as high as 0.204 and 0.024. In contrast, two artists *Beat Streuli* from Switzerland and *Nari Ward* from Jamaica have actually co-participated in several exhibitions, but their coefficients are as low as 0.0208 and 0.0009.

We can find that the previous studies on the research domain put some assumptions implicitly as follows:

Assumption 1 Generally, Web pages are created according to results of two actors' co-participation in the events. Thus, the number of Web pages is assumed to have useful correlation to the number of co-participations.

[5] 17th, 18th and 19th Annual Conferences of the Japan Society of Artificial Intelligence
[6] The Seventh International Conference on Ubiquitous Computing (UbiComp 2005)
[7] Both are names of famous Japanese corporations.

Assumption 2 A community to be extracted as a social network is assumed to be homogeneous.

In the following section, we will introduce our methods that work effectively against these violation of Assumptions.

3.2 Proposed Methods

Relation identification What determines the strength of a tie between two actors? The frequency or degree of relations affects the strength. Multiple relations between the two actors imply a stronger tie. We must identify in which relations two actors are involved to calculate the strength of ties more precisely. In the domain of firms, many relations are published in news reports and on news releases that are distributed on the Web. If the news is given attention by media services or people, many Web pages would describe and comment on the relation. Conversely, if news or tiny companies' relations were not given attention, only a few pages would describe the relations.

As a solution, we add some word or combination of words (called a *relation keyword*) to a search query. Using this strategy, we can efficiently identify the relation between the firms. For example, when we wish to extract lawsuit relations among firms, we add a term *lawsuit*. We issue a query *Matsushita* AND *JustSystem* AND *lawsuit* so that the search engine will return the lawsuit pages that are associated with the two firms, and then we apply text processing to identify partcular relations from retrieved pages. This idea is similar to keyword spices [17], which extend queries for domain-specific web searches. Question answering systems also construct elaborate queries for use with a search engine [18].

Threshold tuning for relation In an inhomogeneous community such as that of artists, even though the relation is observed as weak (with less evidence on the Web), it might be important for the particular actors. For studies undertaken in the social sciences, network questionnaires have been conducted traditionally. Typically, participants are asked "Please name your five closest friends." The response would list the relations that are important for the participant, i.e., subjectively important ones. We articulate the concept and propose more flexile manner to generate a network: We employ two criteria that correspond to objective and subjective importance of relations for actors. We first invent edges by objective criteria using a consistent threshold to invent edges. Then we invent edges using subjective criteria for actors who do not have (a certain number of) edges. The thresholds are optimized so that the target network is extracted as precisely as possible.

4 Social Network Extraction of Firms

We describe the extraction of a firm network in Japan as a case study of relationship identification. A social network of 60 firms, including IT, communication, broadcasting, and electronics firms in Japan, is extracted. The relations are defined among them as alliance and lawsuit relations.

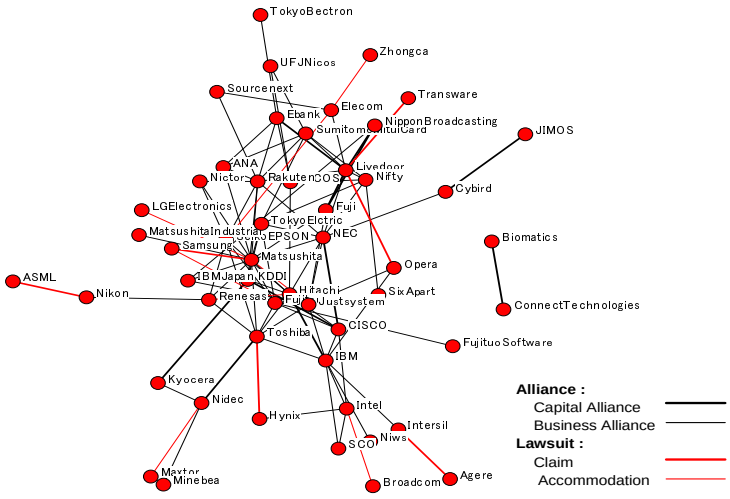


Fig. 1. Social network of 60 firms in Japan.

4.1 System Flow

Two major parts exist in the system: an online module and an offline module.

In the online module, a list of firms is given as an input. Then a query for a search engine is constructed for each pair of firms, adding a relation keyword[8]. We obtain top-ranked five Web documents by submitting each query. A simple pattern-based heuristic is employed to judge the relation:

1. Pick up all sentences that include the names of two firms.
2. Score each sentence by the sum of the scores of relation keywords that are included in the sentence. The pair of the firms gets the maximum score of every sentence as a score.
3. If the score of the pair is greater than a certain threshold[9], i.e., if the two firms seem to have the target relation in high reliability, an edge is invented between the two firms.

In the offline module, relation keywords are obtained beforehand. The intuitive method for finding relation keywords is to select terms that appear often in the target Web pages and do not appear in the remaining pages. For that method, we need to collect annotated Web pages with metadata on alliance/lawsuit relations of the firms as a training corpus. We gathered 456 pages for alliance, and 165 pages for the lawsuit relationship. The training set is collected from news

[8] Actually, we use two relation keywords 提携 (*aline*) AND 株式会社 (*corporate*) and 提携 (*aline*) AND 株式 (*stock*) for alliance relations, and two relation keywords 侵害 (*violate*) AND 訴訟 (*lawsuit*) and 侵害 (*violate*) AND 請求 (*claim*) for lawsuit relations.(Note that the relation keywords are translated from Japanese.) Therefore, four queries are constructed for each pair of firms.

[9] The threshold of the score can be obtained from experience.

Table 1. Precision and Recall of the System.

Target relation	Precision	Recall
Alliance	60.9% (70/115)	62.0% (70/113)
capital alliance	75.0% (9/12)	42.9% (9/21)
business alliance	67.4% (60/89)	60.0% (60/100)
Lawsuit	61.5% (16/26)	100% (16/16)
claim phase	63.6% (14/22)	87.5% (14/16)
accommodation phase	72.7% (8/11)	88.9% (8/9)

pages on NIKKEI NET, and an intellectual property (IP) news site[10]. Then the words (or word combination) which produce the high F-values to discriminate the alliance/lawsuit relationship are selected.

We do not describe the detail due to the space limitation, but we actually categorize alliance relations into two subcategories, capital alliances and business alliances, and lawsuit relations into two subcategories, claim phase and accommodation phase. It is important to distinguish such typical and temporal relations for detailed analyses of social networks. Relations are recognized as non-directional in our current implementation. Please refer to [8] for further details of the algorithm.

4.2 Results and Evaluation

The obtained network for 60 firms in Japan is shown in Fig. 1. Black lines represent alliances (bold ones are capital alliances and thin ones are business alliances) and red lines represent lawsuits (bold ones are in claim phase and thin ones are in the accommodation phase).

The precision and recall of our system is shown in Table 1. For the ${}_{60}C_2 = 1770$ pairs of firms, 113 pairs actually show alliance relations; our system extracted 70 pairs correctly. For subcategories of alliances, there were actually 21 and 100 pairs of capital and business alliances among them, and our system extracted 9 and 60 respectively. Compared to alliances, lawsuit relations have higher recall: We find 16 relations out of 16 actual lawsuit relations. This is probably true because lawsuit relations are described in rather common format using words such as 判決 (judgment), 訴訟 (lawsuit), or 和解 (accommodate).

The extracted relation keywords work effectively: If we do not use relation keywords at all, the average precision/recall is 65.7%/95.0% by query evaluation. If we use the relation keywords, the figures are 87.1%/99.4% using the same number of downloaded documents. By integrating relation keywords to a query, we can search the Web pages more efficiently.

5 Social Network Extraction for Artists

In this section, we describe our algorithm for extracting a social network of artists of contemporary arts: artists at the Yokohama Triennale 2005.

[10] www.nikkei.co.jp and news.braina.com.

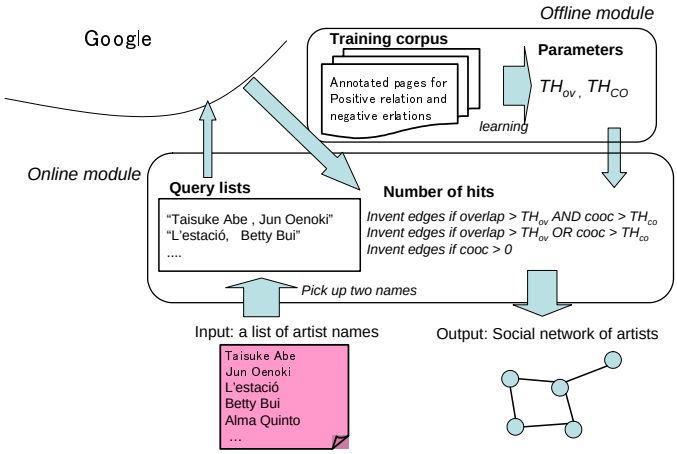


Fig. 2. System flow to extract an artist network.

5.1 System Flow

The whole system is illustrated in Fig. 2. We have an online module to make queries to a search engine and invent edges depending on co-occurrence measures. The offline module tunes the parameters to obtain the precise network.

For each pair of artist names X and Y, we put a query to a search engine. As a result, we obtain two co-occurrence measures: a matching coefficient (denoted as $cooc(X,Y)$) and an overlap coefficient (denoted as $overlap(X,Y)$). Edges are invented using the two thresholds TH_{ov} and TH_{co}. For each actor $x_i \in X_{all}$, the algorithm creates edges by link x_i to those actors who possessing relatively strong relations with x_i.

First, classifies all actors except x_i into following three classes[11]:

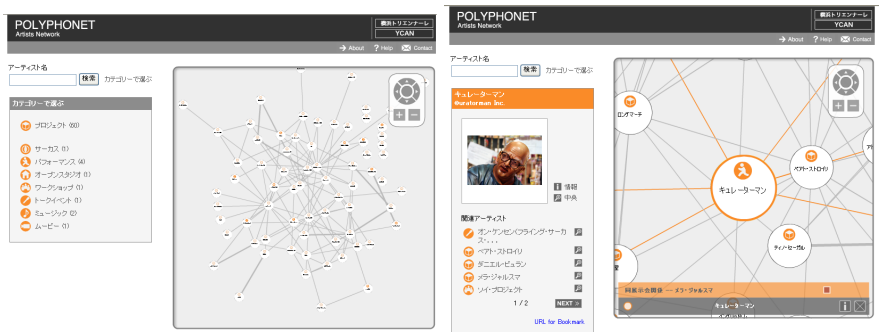
C1 Actors related $x_i \in X_{all}$ with $overlap(X,Y) > TH_{ov}$ and $cooc(X,Y) > TH_{co}$.
C2 Actors related $x_i \in X_{all}$ with $overlap(X,Y) > TH_{ov}$ or $cooc(X,Y) > TH_{co}$.
C3 Actors related $x_i \in X_{all}$ with $overlap(X,Y) > 0$ and $cooc(X,Y) > 0$.

Then, edges are added based on the following steps:

Step 1 Connect every actors from $C1$ with x_i.
Step 2 If the edge number of x_i is less then M, connect actors from $C2$ until it reaches M.
Step 3 If the edge number of x_i is less than $N (N \leq M)$, connect the actors in $C3$ to x_i until edge number of x_i reaches N or all the actors $C3$ are connected.

This procedure alleviates the problem of some nodes having too many edges and some nodes being isolated. It enables more exhaustive extraction for every

[11] The actors in all of three classes are sorted by $overlap(x_i, Y)$. Those actors do not belong to any classes are ignored in the next step.



(a) The whole network. (b) Centering artist *Curatorman*.

Fig. 3. System Interface for Yokohama Triennale 2005.

node compared to the previous method, although it sometimes yields low precision. For that reason, we must find the appropriate parameters for TH_{ov}, TH_{co}, M and N.

5.2 Evaluation

We tune the threshold TH_{ov} and TH_{co} as well as M and N in the offline phase. In the Triennale case, we use 1000 pairs of artists as training data, consisting of 146 positive examples and 854 negative examples.[12] The parameters are tuned so that the F-value of the examples is maximized. Using the evaluation set, the proposed algorithm (executing Step 1-3) produces the F-value 0.55, which is more than 0.05 points better than the existing methods (executing only Step 1). The greatest difference is that the proposed algorithm produces high recall while maintaining modest precision. It is important when the purpose is to promote navigation and communication using a social network.

5.3 Navigation Sites for Yokohama Triennale

Our system is in operation on the official support site for Yokohama Triennale 2005 to provide an overview of the artists (133 artists with 71 projects) and informational navigation for users. At exhibitions, it is usual that participants enjoy and evaluate each work separately. However, our concept is that if participants know the background and relations of the artists, they might enjoy the event more. For that purpose, the system provides relations of artists and that evidence for users.

The system interface is shown in Fig. 3. It is implemented using Flash and interactive navigation is realized. The system provides a retrieval function. The information about the artist is shown on the left side if a user clicks a node. In addition, the edges from the nodes are highlighted in the right-side network.

[12] 1000 pairs are not large considering that the potential number of edges for 133 artists is 8778.

Table 2. Centrality

(a) Eigenvector centrality.

Rank	Name	Value
1	Matsushita	0.366
2	Hitachi	0.351
3	NEC	0.289
4	Fujitsu	0.275
5	Toshiba	0.263
6	Rakuten	0.257
7	Just System	0.241
8	KDDI	0.208
9	Tokyo Electric	0.207
10	Seiko Epson	0.204

(b) Betweenness centrality.

Rank	Name	Value
1	Matsushita	168.981
2	IBM	149.192
3	NEC	144.675
4	Hitachi	136.978
5	Toshiba	113.239
6	Rakuten	109.887
7	Just System	77.175
8	Livedoor	74.141
9	CISCO	64.558
10	Fujitsu	56.081

The user can proceed to view the neighboring artists' information sequentially, and can also jump to the Web pages that show evidence of the relation. We encourage the reader to visit the website at www.tricosup.org.

6 Application and Discussion

We have described the improved methods to extract social networks of various entities, particularly of firms and artists. The obtained network can be useful in the context of the Semantic Web. For example, we can use a social network of artists for detecting COI among artists when they make evaluations and comments on others' work on the Web. Furthermore, we can find firm clusters and characterize a firm by its groups.

We present a prototypical example of applications using a social network of firms. We calculate the centrality for each firm on the extracted social network (on alliance). Table 2(a) shows the top ten firms by eigenvector centrality. These firms have remained large and reliable corporations in Japan for decades. Table 2(b) shows the top ten by betweenness centrality. Interestingly, IBM, Livedoor and Cisco are on the list. These firms might bridge two or more clusters of firms: IBM and Cisco are United States' firms and form alliances with firms in multiple clusters; Livedoor is famous for its aggressive M & A strategy in Japan. There seem to be many potential applications that can make use of social networks in the Semantic Web.

7 Conclusion

This paper describes extraction methods of various social networks from the Web. To date, numerous studies of social networks have addressed the researcher domain. It is an important test-bed; however, the next step must be taken to leave the domain of researchers. This paper steps further to show that researcher networks might be an easy domain for social network extraction from the Web. The main points of our methods can be summarized as (i) to use a search engine more efficiently and integrate text processing, and (ii) to tune network generation according to relational data depending on the target domain.

We believe that a network point of view is important for knowledge integration and articulation as well as for (lightweight) ontology emergence. We intend to develop general-purpose (social) network extraction in the future. The intersection of social network and ontology emergence might become a fertile ground for Semantic Web research.

References

1. B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, A. Sheth, I. Arpinar, L. Ding, P. Kolari, A. Joshi, and Tim Finin. Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection. In *Proc. WWW2006*, 2006.
2. R. Bekkerman and A. McCallum. Disambiguating web appearances of people in a social network. In *Proc. WWW 2005*, 2005.
3. D. Bollegara, Y. Matsuo, and M. Ishizuka. Disambiguating personal names on the web using automatically extracted key phrases. In *Proc. ECAI 2006*, 2006.
4. A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *CEAS-1*, 2004.
5. T. Finin, L. Ding, and L. Zou. Social networking on the semantic web. *The Learning Organization*, 2005.
6. S. Ghita, W. Nejdl, and R. Paiu. Semantically rich recommendations in social networks for sharing, exchanging and ranking semantic context. In *Proc. ISWC05*, 2005.
7. J. Golbeck and B. Parsia. Trust network-based filtering of aggregated claims. *International Journal of Metadata, Semantics and Ontologies*, 2006.
8. Y.Z. Jin, Y. Matsuo, and M. Ishizuka. Extracting inter-business relationship from world wide web. *Journal of Japanese Society for Artificial Intelligence*, 2006. Submitted. Temporary available at http:/ymatsuo.com/papers/jsai06Jin.pdf.
9. H. Kautz, B. Selman, and M. Shah. The hidden Web. *AI magazine*, Vol. 18, No. 2, pp. 27–35, 1997.
10. X. Li, P. Morie, and D. Roth. Semantic integration in text: From ambiguous names to identifiable entities. *AI Magazine Spring*, pp. 45–68, 2005.
11. P. Massa and P. Avesani. Controversial users demand local trust metrics: an experimental study on epinions.com community. In *Proc. AAAI-05*, 2005.
12. Y. Matsuo, M. Hamasaki, H. Takeda, J. Mori, D. Bollegala, Y. Nakamura, T. Nishimura, K. Hasida, and M. Ishizuka. Spinning multiple social networks for semantic web. In *Proc. AAAI-06*, 2006.
13. Y. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hasida, and M. Ishizuka. POLYPHONET: An advanced social network extraction system. In *Proc. WWW 2006*, 2006.
14. P. Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, Vol. 3, No. 2, 2005.
15. P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proc. ISWC2005*, 2005.
16. J. Mori, M. Ishizuka, T. Sugiyama, and Y. Matsuo. Real-world oriented information sharing using social networks. In *Proc. ACM GROUP'05*, 2005.
17. S. Oyama, T. Kokubo, and T. Ishida. Domain-specific web search with keyword spices. *IEEE TKDE*, Vol. 16, No. 1, pp. 17–27, 2004.
18. G. Ramakrishnan, S. Chakrabarti, D. Paranjpe, and P. Bhattacharyya. Is question answering an acquired skill? In *Proc. WWW2004*, 2004.