Towards Discovering Organizational Structure from Email Corpus

Ding Zhou[1] Yang Song[1] Hongyuan Zha[1,2]
Dept. of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16801, U.S.A.

Ya Zhang[2]
Information Sciences and Technology[2]
The Pennsylvania State University
University Park, PA 16801, U.S.A.

Abstract

Email logs people's communication history which provides valuable information regarding the infrastructure of an organization. In this paper, a two-phase framework is introduced to attack the problem of leadership discovery in an organization based on email communication history among the employees. Two heuristic metrics are proposed for evaluating pair-wise leadership factors among a group of employees. We also address several issues in discovering the organization's structure through mining leadership graph constructed from the leadership factors. Experimental studies are carried out by applying the framework to Enron email corpus.

1. Introduction

Social Network (SN) has been recognized as one of the most convenient, direct ways to represent the relationships among a group of individuals. In a typical *SN*, each node represents a person, and two nodes share an edge if they are related, based on certain metric.

Research on *Social Network Analysis (SNA)* spans the fields of sociology, psychology, management science, and computer science. Issues covered by *SNA* in computer science include but not limited to cluster analysis [7, 16, 3], node centrality measurements [6], network topology [1, 2]. The cluster analysis explores community structure of an *SN* based on different metrics such as betweenness centrality [16]. Node centrality measures the importance or prominence of a node in the network according to its "location". The network topology studies structures of network such as structural equivalence and structural holes. Some recent research proposed to extract the personal information from both the email and personal web site, from which cluster of individuals are formed[4].

While the existence of leadership in an *SN* is widely recognized, there have been very few literatures on the discovery of it. The concept of leadership in *SN* has been studies for decades and is still of interests [9] [17]. Leadership defines the opinion leaders as the brokers who carry information across the social boundaries between groups. This heuristic identification of leaders, however, suffers from lack of pair-wise leadership evaluation. As a result, the leadership between two individuals is hard to quantify.

Another problem with existing *SNA* research is the use of un-directed network, with which certain relationships between individual might be lost. Consider the *SN* of employees in a corporation, where hierarchical structure embeds. At least three types relationships exist among individuals: colleague of, supervisee of, and supervisor of. The latter two relationships, which are not reversible, calls for representations more than just an un-directed edge. A directed weighted graph, where the direction indicates the type of relationships and the weight indicates the strength of the relationship, may release the problem and better reflect such *SN* with organizational information.

In this paper, we propose to discover the organization structure as a directed weighted graph. The organizational structure is generated in two steps. We first evaluate the leadership factors between pair of individuals and construct the directed *SN*. Then the organizational structure is discovered from the *SN* we obtain. We propose two metrics to facilitate the *SN* construction from email corpora.

We study the *SN* extracted from emails because emails embodies valuable information regarding knowledge exchange and the infrastructure within a social network [15, 16]. Various formats of resources, such as message boards [13], emails [16] or web sites [4], have been utilized for the discovery of social relationships. Recently, the availability of Enron email corpus has been considered as a touchstone for exploring emails as the resource in studying social network.

2. Discovering Organizational Community Structure from Email Corpus

We describe a two-phase framework for the discovery of organizational community structure within an email corpus.

In the first phase, we evaluate the pair-wise leadership information between every two email users. In the second phase, the organizational structure embedded in the society of email users is discovered.

2.1. Pair-wise leadership Evaluation

In an information society, the problem of pair-wise leadership evaluation is formalized as: given an organization of people, say U, let u_i and u_j be two individuals, determine the leadership factor (LF) between them. Define $LF_{ij} = \xi(u_i, u_j, \Omega_U)$,where ξ is a function that represents the degree of u_i's leadership over u_j. Ω_U is any reliable resource used as indicator for such leadership information in U. In this paper, we have narrowed to only using the email communications among users as the indicator resource. The key issue is to design a function that properly evaluates the leadership matrix $\{LF_{i,j}\}$.

In this paper, we concern ourselves with two heuristic methods (LF^1 and LF^2) founded at the use of email communication history. One method use the sender-receiver imbalance and the other builds on the group information inferred in email group lists.

2.1.1 LF^1 : imbalance between sender and receiver

One interesting characteristic of a social network is the *degree disparity* [10]. This condition arises when the objects in the network have widely different distributions of degree, which is defined as the number of edges to/from other objects. Such characteristic can be utilized for evaluation of the pair-wise relationship.

In an email social network graph constructed based on communication frequencies on pairs of users, the vertices represent email users and edges present between two users who corresponded through email frequently enough [16].

For our first scenario LF^1, we extend the graph by weighting the directed edge E_{ij}. The weight on E_{ij} is determined by the number of messages that user u_i sends to u_j. We propose a metric of LF based on the imbalance between sender and receiver. In our first metric LF^1, the LF_{ij} between two users is evaluated as follows:

$$LF^1_{ij} = \frac{P_{j,i}}{P_{i,j} + P_{j,i}}, \tag{1}$$

where $P_{i,j}$ and $P_{j,i}$ are the percentages of emails that u_i sends to and receives from u_j [1].

The idea behind LF^1 is tentative: given two email users in an organization, the one at leading position is more likely to get filled with send-in messages while is less willing to respond to his/her subordinates. Such notion parallels with one web page ranking principle recognized in search engine application. When ranking the retrieved web pages for a user query on a Web search engine, the web pages with more other pages linking to is considered more authoritative and thus is ranked higher [11, 5]. Similarly in an email social network, the more messages u_i receives from u_j as such as a link in the social network, the higher likely it is that u_i is at an authoritative position that leads u_j.

Consider a pair of users u_i and u_j, the leadership relation between them may be in either of the two cases: (1) u_i leads(or is leaded by) u_j by the factor of $LF^1_{ij}(LF^1_{ji})$; (2) there is no direct leadership relation between u_i and u_j.

2.1.2 LF^2 : inference from email group list

Another method to measure the LF_{ij} is based on the inference from the group lists in an email corpus. We define the $group\ list$ of an email message as the list of users that are accessible to this email. Typically, a group list contains the users in fields of $From$:, To :, CC :, BCC :, etc.

Obviously, an email group list infers the set of users that are associated by this email message. Given a set of email messages, each of which indicates a small group and corresponding members, the $support$ of a group list is counted as the frequency that such list presents in the message set. Conceptual groups, or clustering of users, can easily be discovered by simply retrieving the most frequently presented group lists. The higher $support$ a group list has, the more likely it is that the set of users are closely related.

In our second metric LF^2, we propose to score LF^2_{ij} using the conditional probability that u_i co-occurs given u_j is present in a list:

$$LF^2_{ij} = P(u_i|u_j), \tag{2}$$

which re-formulates as:

$$P(u_i|u_j) = \frac{support(u_i \cap u_j)}{support(u_j)} \tag{3}$$

Consider a simple organization of four persons A, B, C and D. A leads B, C, D which are on the same level. We assume that they interact primarily based on emails. Table 1 summarizes the group lists and corresponding $support$s.

From Table 1, we have the the following LF^2: $LF^2_{A,B} = (10+6)/(10+6+2) = 0.889$ and $LF^2_{A,C} = (10+7)/(10+7+2) = 0.895$. Similarly, $LF^2_{B,C} = 0.105$, $LF^2_{B,A} = 0.571$, $LF^2_{C,A} = 0.607$, $LF^2_{C,B} = 0.111$. Compared with the original organization, we can see that when there is a leadership of X over Y,$LF^2_{X,Y}$ is high while $LF^2_{Y,X}$ is low.

The computation of equation 3, however, is not as straightforward as it looks like. In a practical email corpus,

[1]There can be cases when $P_{i,j} + P_{j,i} = 0$. However, it is trivial to filter these "non-active" users before processing.

group list	*support*
A, B, C, D	10
A, B	6
B, C	2
A, C	7
A, D	5

Table 1. Group lists in a simple organization

where there are a large number of emails, the brute-force statistical computation results in much overhead. For efficient approximate measure of the our proposed LF^2, we introduce an approximate algorithm $LF^2measure$ in Figure 1.

Input: GL, U, $support$
Output: $\{LF^2_{ij}\}$
(1) $MFGL \leftarrow$ fpMine$(GL, support)$
(2) /* $MFGL$ is set of maximal frequent group lists */
(3) for each $u_i \in U$
(4) for each $u_j \in U$
(5) $c_{u_i,u_j} = \sum_{\{L,p\} \in MFGL, u_i \in L, u_j \in L} p$
(6) $c_{u_j} = \sum_{\{L,q\} \in MFGL, u_j \in L} q$
(7) $LF^2_{ij} = \frac{c_{u_i,u_j}}{c_{u_j}}$

Figure 1. Algorithm $LF^2measure$

The algorithm $LF^2measure$ takes GL, the set of group lists, U, the user set and the parameter $support$. It invokes any frequent pattern mining algorithm, say $fpGrowth$ [8], using the $support$. At line (1) in Figure 1, the maximal frequent group lists and corresponding frequencies are computed and kept in $MFGL$. A list L is maximal frequent if it shows up more than $support$ times in GL and there is no other frequent group list in $MFGL$ that contains L.

2.2. Discovering Organizational Structure from Pair-wise Leadership

In the former section, we have introduced the first step towards discovering the organization structure from an email corpus. In this section, we describe the second step which discovers the organizational structure of an information society. The second step is based on the pair-wise LF obtained from the first step.

Let us start with a simple example: suppose the pair-wise LF evaluated from the fact Table 1 is as shown in Table 2 (only LF^2 is presented for simplicity). The ij entry of the matrix $\{LF_{ij}\}$ denotes u_i's leadership over u_j. On discovering the organizational structure within this information society, our first question is: *given a person u_i, who is most likely to be the leader of u_i, i.e. the identification of leader?*

	A	B	C	D
A	/	0.889	0.895	1.000
B	0.571	/	0.105	0.667
C	0.607	0.111	/	1.000
D	0.574	0.556	0.526	/

Table 2. Pair-wise LF^2 Matrix

2.2.1 Identification of Leader

Look at the $\{LF_{ij}\}$ matrix in Table 2. Suppose we want to determine B's leader. Intuitively, we compare LF_{AB}, LF_{CB} and LF_{DB} to find out that A has the largest LF over B. However, we can not conclude A is B's leader simply by the fact that $LF_{AB} = max\{LF_{iB}|LF_{iB} \in \{LF_{ij}\}\}$. Consider the case when the $\{LF_{ij}\}$ among users are as illustrated in Figure 2. Both LF_{PQ} and LF_{QP} are very high but close. LF_{RP} is much larger than LF_{PR} while LF_{RP} is smaller than LF_{QP}[2]. In such case, selecting the leader of u_j by maximizing LF_{ij} is no longer reliable, when we select Q as P's singular leader[3].

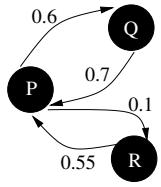


Figure 2. Identification of Leader for P

We propose to classify u_i as leader for u_j based on such criteria that: (1) u_i is most likely to be u_j's leader and (2) u_j is least likely to lead u_i, i.e. we search for such an i that maximizes LF_{ij} and minimizes LF_{ji} at meantime.

2.2.2 Discovery of Organizational Structure

Once the identification of leader is feasible, the discovery of organizational structure becomes straightforward. We propose a greedy algorithm: from all the pair-wise leadership factors, we choose the pair of users u_i and u_j with maximal $(LF_{ij} - LF_{ji})$, add the directed edge $\overrightarrow{E_{ij}}$ to the social network without embedding cycles into the graph. We avoid a

[2]Obviously such case does not exist for LF^1 measure where $LF_{ij} + LF_{ji} = 1$, but there are LF measurements that violates this.

[3]Note that there are some organizations that allows multiple leadership, i.e. u_i can be leaded by both u_j and u_k. This can be feasibly discovered, however, by extending our framework, which we do not cover in this paper due to the limited space.

cyclic organizational graph based on the notion that in most organizations the leadership relation is acylic. Iteratively, we follow the procedure until there are no users left.

Let $\{LF_{ij}\}$ be the matrix of leadership factor, U be the user group and T be the organizational tree as output. Formally, our proposed algorithm for Discovering Organizational Structure(DOS) can be defined as in Figure 3.

Input: $\{LF_{ij}\}, U$
Output: T
(1) $T \leftarrow \emptyset$
(2) let $\delta_{i,j} = LF_{ij} - LF_{ji}$
(3) $\overrightarrow{E} \leftarrow \{(i,j,k) | \delta_{i,j} \geq \delta_{i',j'}$ if $k < k'\}$
(4) /* (i,j,k) denotes a directed edge $\overrightarrow{E_{ij}}$
(5) is located kth in the sorted list of edges $\overrightarrow{E}$ */
(6) for each edge $(i,j,k) \in \overrightarrow{E}$
(7) $m = max\{j | u_j \in U, T \cup (i,j)$ is acylic tree$\}$
(8) $T = T \cup (i,m)$
(9) $R = \{u_k | u_k \in U, u_k$ has no leader$\}$
(10) /* R contains the roots of T */

Figure 3. Algorithm DOS

3 Experiments

We present experimental results of our framework with the Enron email corpus. The Enron email was made public by the Federal Energy Regulatory Commission during the investigation. William Cohen from CMU prepared the dataset and has it published on the web for researchers [12] This version contains 517,431 emails from 151 users distributed in 3500 folders.

We created a MySql database to support efficient mining on the large scale of emails. Our database schema extended [14] by adding another table of email flows among Enron employees. Non-Enron employees were filtered because they were not helpful in identifying the social organizational structure within the corporation.

3.1. Leadership Factor Evaluation

3.1.1 LF^1 v.s. LF^2

We present experiments with our two metrics in LF evaluations. In Figure 4, the $\{LF^1_{ij}\}$ measure using sender-receiver imbalance is presented. We illustrate the LF^1 using gray intensity of a cell in the picture. There are four kinds of cells in terms of gray intensities in Figure 4. We label the high LF cell with high intensity. While for cells where no leadership exists, we mark it white.

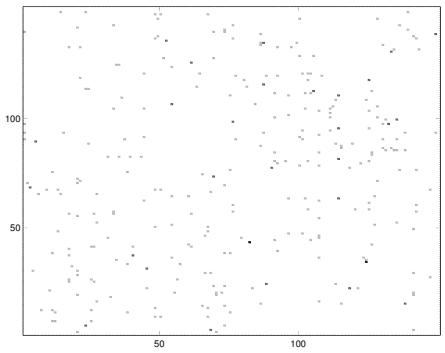


Figure 4. Visualization of the LF^1_{ij} Matrix

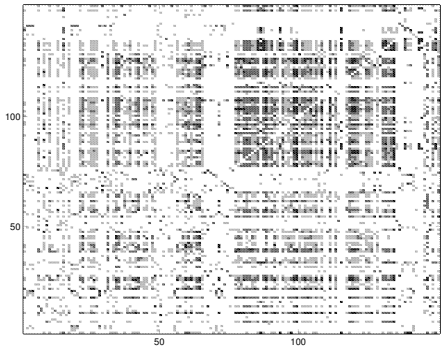


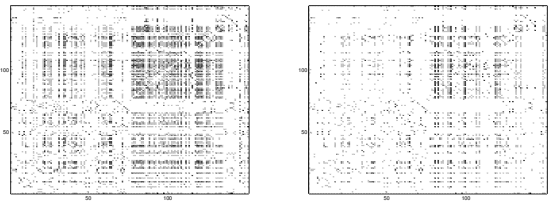
Figure 5. LF^2_{ij} Matrix, $support$=1

We can see from the Figure 4 that only a small number of the user-pairs indicate strong LF^1 while most others either have low LF or remain unrelated. This implies that the utilization of sender/receiver imbalance for leadership measuring might suffer from the loss of pair-wise relationship.

Compared with $\{LF^1_{ij}\}$, the $\{LF^2_{ij}\}$ matrix seems much packed. The $\{LF^2_{ij}\}$ discovered with $support = 1$ is presented in Figure 5. Meanwhile, in Figure 6, the $\{LF^2_{ij}\}$ matrixes under different $supports$ are illustrated. We can see from these figures that the pair-wise leadership relation become looser as the $support$ sets higher. It is because that when the $support$ gets high, some less frequent group lists are filtered, which results in the loss of grouping information while deliberating leadership. We also found from our experiments that the acquirement of $\{LF^2_{ij}\}$ requires higher computational cost. For efficiency concern, we present our experimental results using the approximation method introduced in Section 2.1.2.

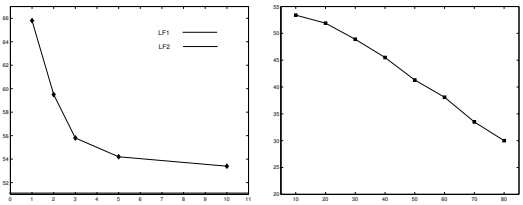
3.1.2 Efficiencies

In this subsection, we examine the efficiency issue for LF measurement considering the large number of emails in a



(a) $support$=5 (b) $support$=20

Figure 6. LF_{ij}^2 **Matrix with** $support = 5, 10$



(a) small $support$ (b) large $support$

Figure 7. Runtime for LF^2 **w.r.t. different** $supports$

normal email corpus. As aforementioned, the LF^1 requires low computational cost while this criterion might suffer from the loss of pair-wise relationship. By contrast, the LF^2 yields stronger leadership information but the computation costs much.

Figure 7(a) shows the runtime for both LF^1 and LF^2 with small $support$. The efficiency advantage of LF^1 is obvious. In Figure 7(a) through Figure 7(b), the scalability of LF^2 to the setting of $support$ is presented. We can see that except for the loosened pair-wise leadership matrix, the use of low $support$ may benefits, in terms of efficiency. Obviously, when the $support$ is set to 1, the $LF^2measure$ degrades to a brute-force method but is yet most precise.

3.2. Organizational Structure Discovery

In this subsection, we show the organizational structure discovered from the pair-wise LF matrix obtained from the first step of our framework. There can be various approaches for representation of the organizational structure in an information society. One widely accepted way is the tree-like organization chart(OC), the other way is the directed acylic graph(DAG).

Comparing OC and DAG representations, one difference is in their definition of leadership singularity. In a OC representation, there can be no more than one leader for an entity while it is not the case in DAG. Another difference comes from their maintenance of information completeness. A DAG keeps more leadership information by allowing one entity leaded by multiple entities. In addition, we will see they also differs in scalability. Due to the limit in space, we choose to present the OC representation of the organizational structure in consistence with leadership singularity.

We ran our DOS algorithm on both $\{LF_{ij}^1\}$ and $\{LF_{ij}^2\}$ measurements. The discovered OCs are presented in Figure 8(a) and Figure 8(b). For visualization and privacy concerns, we tag the 151 Enron employee in the with their $employeeID$, which ranges from 1 to 151. We also highlight the Enron employee at certain positions such as $president$ or $director$ with different colors. The $president$ employee is colored blue and $director$ red.

The OC based on LF^1, in Figure 8(a), consists of three trees. It is not surprising to see that all the three trees are rooted at individuals whose positions are not high-level. Note that the OC we discovered is not equivalent to the hierarchy by positions. It is often the someone, who is not at the top position, that coordinates the entire Corp. In Figure 8(b), the OC seems more organized and the levels of hierarchy is more reasonable. Also the OC tree in Figure 8(b) is better balanced. For OC based on LF^1, we discovered twelve levels and for LF^2 we found five levels.

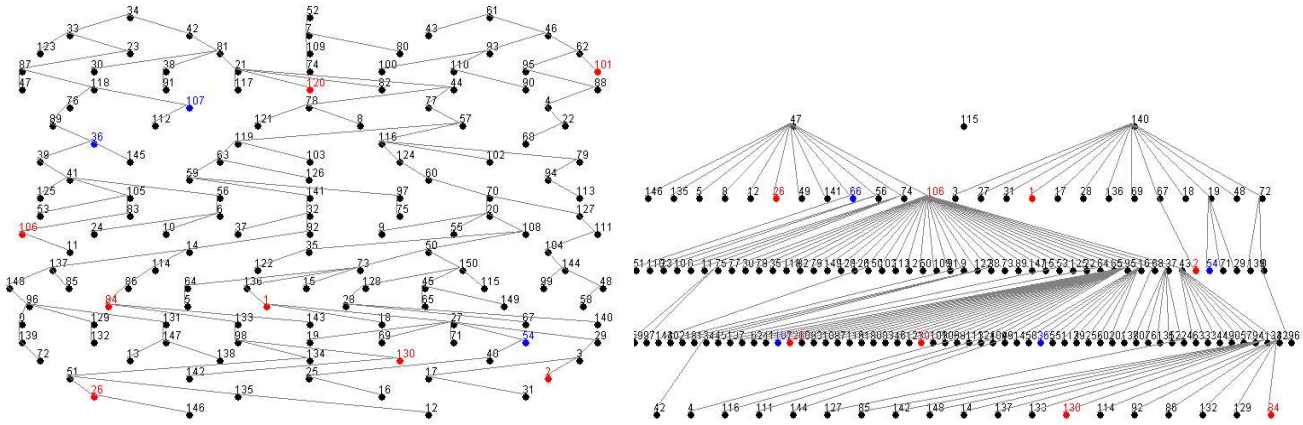
For OC based on LF^2, we present the levels and corresponding numbers of entities per level, in Table 3. We have also posted the Enron $employeeID$ and corresponding positions at "http://www.cse.psu.edu/cmla/download/positionList.xls". We obtained such former Enron employee status report from [14]. Due to privacy concerns, however, some persons or their positions are missing in the list.

levels:	1st	2nd	3rd	4th	5th
entity #:	3	25	47	56	19

Table 3. Entities per level in OC **from** LF^2

4 Conclusions

In this paper, we introduced a two-phase framework for discovering organizational structure from an information society. We proposed to use email as the resource for the discovery of such social leadership relations. Two heuristic metrics were proposed to evaluate pair-wise leadership factors among a group of individuals. Experimental results showed that our framework works properly on Enron Email Dataset. Meanwhile, we addressed several issues in recon-



(a) LF^1-based OC (b) LF^2-based OC, $support$=5

Figure 8. Runtime for LF^2 w.r.t. different $supports$

structing the organizational structure of an information society that brews plentiful extensions of our work.

References

[1] G. Ahujia. Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative Science Quarterly*, pages 425–455, 2000.

[2] R. S. Burt. Relational equilibrium in a social topology. *Journal of Mathematical Sociological*, 1979.

[3] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[4] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *First Conference on Email and Anti-Spam (CEAS 2004)*, 2004.

[5] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon. Link analysis: Hubs and authorities on the world. *Journals of the Society for Industrial and Applied Mathematics*, 2001.

[6] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, pages 215–239, 1979.

[7] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, pages 7821–7826, 2002.

[8] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 1–12, New York, NY, USA, 2000. ACM Press.

[9] D. Jensen and J. Neville. The social capital of opinion leaders. *Annal of the American Academy of Political and Social Science*, 1999.

[10] D. Jensen and J. Neville. Data mining in social networks. *Proc. National Academy of Sciences Symposium on Dynamic Social Network Analysis*, 2002.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[12] B. Klimt and Y. Yang. Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS 2004)*, 2004.

[13] N. Matsumura, D. E. Goldberg, and X. Llora. Mining directed social network from message board. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 1092–1093, New York, NY, USA, 2005. ACM Press.

[14] J. Shetty and J. Adibi. The enron email dataset database schema and brief statistical report. In *Technical Report*, 2004.

[15] L. Sproull and S. Kiesler. Reducing social context cues: electronic mail in organizational communication. *Manage. Sci.*, 32(11):1492–1512, 1986.

[16] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman. Email as spectroscopy: automated discovery of community structure within organizations. *Communities and technologies*, pages 81–96, 2003.

[17] S. Wasserman and K. Faust. Social network analysis. *Cambridge University Press*, 1994.