Online Supplement to "A Network Structural Approach to the Link Prediction Problem"

Chungmok Lee
IBM Research–Ireland, Damastown Industrial Park, Mulhuddart, Dublin 15, Ireland / RUTCOR, Rutgers University, Piscataway, New Jersey 08854, USA, chungmok@gmail.com,

Minh Pham
RUTCOR, Rutgers University, Piscataway, New Jersey 08854, USA, ptuanminh@gmail.com,

Myong K. Jeong
RUTCOR, Rutgers University, Piscataway, New Jersey 08854, USA, mjeong@rci.rutgers.edu,

Dohyun Kim
Department of Industrial & Management Engineering, Myongji University, 116 Myongji-ro, Cheoin-gu, Yongin, Gyunggi-do, South Korea, norman.kim@gmail.com,

Dennis K. J. Lin
Department of Statistics, Eberly College of Sciences, Pennsylvania State University, University Park, Pennsylvania 16802, USA, dennislin@psu.edu,

Wanpracha Art Chavalitwongse
Departments of Industrial & Systems Engineering and Radiology, University of Washington, Seattle, Washington 98195, USA, artchao@uw.edu,

Appendix A: Proof of Theorem 1

The problem is clearly in NP because the maximum weight b-matching problem can be solved in a polynomial time (Anstee 1987). For showing NP-completeness, we use a reduction from SAT. For any instance of SAT, let $U = \{u_1, u_2, \ldots, u_p\}$ and $C = \{c_1, c_2, \ldots, c_q\}$ denote the set of variables and the set of clauses, respectively. We construct graph $G(V, E)$ and parameters as follows.

$$
\begin{aligned}
V &= V_c \cup V_o \cup V_r \cup V_t \cup \left\{ \bigcup_{i=1,\ldots,p} V_u^i \right\}, \\
V_c &= \{v_1^c, v_2^c, \ldots, v_q^c\}, \\
V_o &= \{v_{1,1}^o, v_{1,2}^o, \ldots v_{1,q}^o, v_{2,1}^o, v_{2,2}^o, \ldots v_{2,q}^o, \ldots, v_{q,1}^o, v_{q,2}^o, \ldots v_{q,q}^o\}, \\
V_u^1 &= \{v_1^u, v_1^{\neg u}\}, \\
V_u^2 &= \{v_2^u, v_2^{\neg u}\}, \\
&\vdots \\
V_u^p &= \{v_p^u, v_p^{\neg u}\}, \\
V_r &= \{v_1^r, v_2^r, \ldots, v_p^r\}, \\
V_t &= \{v_{1,1}^t, v_{1,2}^t, \ldots v_{1,q}^t, v_{2,1}^t, v_{2,2}^t, \ldots v_{2,q}^t, \ldots, v_{p,1}^t, v_{p,2}^t, \ldots v_{p,q}^t\},
\end{aligned}
$$

$$
\begin{aligned}
E &= E_{c,o} \cup E_{r,t} \cup E_{u,r} \cup E_{c,u} \cup E_{c,\neg u},\\
E_{c,o} &= \{\{v_i^c, v_{i,j}^o\} \mid i = 1,\ldots,q, j = 1,\ldots,q\},\\
E_{r,t} &= \{\{v_i^r, v_{i,j}^t\} \mid i = 1,\ldots,p, j = 1,\ldots,q\},\\
E_{u,r} &= \{\{v_i^u, v_i^r\} \mid i = 1,\ldots,p\} \cup \{\{v_i^{\neg u}, v_i^r\} \mid i = 1,\ldots,p\},\\
E_{c,u} &= \{\{v_i^c, v_j^u\} \mid j = 1,\ldots,p, i = 1,\ldots,q, \text{and clause } c_i \text{ contains variable } u_j\},\\
E_{c,\neg u} &= \{\{v_i^c, v_j^{\neg u}\} \mid j = 1,\ldots,p, i = 1,\ldots,q, \text{and clause } c_i \text{ contains variable } \neg u_j\},\\
s_e &= \begin{cases} 1, & \text{if } e \in E_{c,u} \cup E_{c,\neg u} \\ M, & \text{if } e \in E_{c,o} \cup E_{r,t} \\ N, & \text{if } e \in E_{u,r} \end{cases}, \ \forall e \in E,\\
\hat{b}_i &= \begin{cases} q+1, & \text{if } i \in V_c \cup V_o \cup V_r \cup V_t \cup \{v_1^u, v_2^u, \ldots, v_p^u\} \\ 0, & \text{if } i \in \{v_1^{\neg u}, v_2^{\neg u}, \ldots, v_p^{\neg u}\} \end{cases}, \ \forall i \in V,\\
L &= (q^2 + pq)M + pN + q,
\end{aligned}
$$

where $N := 2pq + 1$ and $M := 2pN + 1$. Note that there can be at most $2pq$ edges with edge score 1 (i.e., $s_e = 1$) while exactly $q^2 + pq$ edges and $2p$ edges have edge scores M and N, respectively.

Figure A.1 illustrates an example of the reduction.

Claim A.1. *For any permutation P, $F(s, P\hat{b}) \le L$ holds.*

Assume that for some permutation $\hat{P}$ we have $F(s, \hat{P}\hat{b}) > L$ with matching solution $\hat{x}$. It is obvious that $\hat{x}_e = 1$ for all $e \in E_{c,o} \cup E_{r,t}$ so that $(\hat{P}\hat{b})_i = q + 1$ for all $i \in V_c \cup V_r$, that implies $\sum_{e \in E_{u,r}} \hat{x}_e = q$. Thus, $F(s, \hat{P}\hat{b}) = (q^2 + pq)M + pN + \sum_{e \in E_{c,u} \cup E_{c,\neg u}} \hat{x}_e = L - q + \sum_{e \in E_{c,u} \cup E_{c,\neg u}} \hat{x}_e$. By assumption, we have $\sum_{e \in E_{c,u} \cup E_{c,\neg u}} \hat{x}_e > q$ that means there is $i^* \in V_c$ such that $(\hat{P}\hat{b})_{i^*} > q + 1$ which derives a contradiction.

Let $\mathbb{P}$ be the set of all permutation matrices. And let $\hat{\mathbb{P}} := \{P \in \mathbb{P} \mid (P\hat{b})_i = q + 1, \text{ for all } i \in V_o \cup V_c \cup V_r \cup V_t, \text{ and } (P\hat{b})_{v_j^u} + (P\hat{b})_{v_j^{\neg u}} = q + 1 \text{ for all } j = 1,\ldots,p\}$, i.e., $\hat{\mathbb{P}}$ is a set of perturbations that all nodes in $V_o \cup V_c \cup V_r \cup V_t$ have degree constraints of $q+1$ and exactly one of two nodes in V_u^j has node degree constraint of $q + 1$.

Claim A.2. *$P \in \hat{\mathbb{P}}$ if and only if $F(s, P\hat{b}) \ge L - q$.*

The sufficient condition is obvious. For showing the necessary condition, assume that there exists $P^* \in \mathbb{P} \backslash \hat{\mathbb{P}}$ such that $F(s, P^*\hat{b}) \ge L - q$. It is clear that $(P^*\hat{b})_i = q + 1$ for all $i \in V_o \cup V_c \cup V_r \cup V_t$ (otherwise $F(s, P^*\hat{b}) \le L - M$). Since $P^* \in \mathbb{P} \backslash \hat{\mathbb{P}}$ there is some i^* such that $(P^*\hat{b})_{v_{i^*}^u} + (P^*\hat{b})_{v_{i^*}^{\neg u}} = 0$ that implies $F(s, P^*\hat{b}) \le L - N$ which derives a contradiction.

Claim A.3. *If $P \in \hat{\mathbb{P}}$, $F(s, P\hat{b}) = L - q + \sum_{e \in E_{c,u} \cup E_{c,\neg u}} x_e^*$, where x^* is a solution of b-matching problem $F(s, P\hat{b})$.*

This is clear by Claim A.2.

For any permutation $P \in \hat{\mathbb{P}}$, define truth assignment $T_P : U \to \{true, false\}$ as follows.

$$
T_P(i) = \begin{cases} true, & \text{if } (P\hat{b})_{v_i^u} = q + 1 \text{ and } (P\hat{b})_{v_i^{\neg u}} = 0; \\ false, & \text{if } (P\hat{b})_{v_i^u} = 0 \text{ and } (P\hat{b})_{v_i^{\neg u}} = q + 1. \end{cases}, \text{ for all } i = 1,\ldots,p.
$$

We now claim that C is satisfiable if and only if there is a permutation matrix $\tilde{P}$ such that $F(s, \tilde{P}\hat{b}) = L$.

For a sufficient condition, assume that C is satisfiable for truth assignment T^*. We consider a permutation matrix $P^* \in \hat{\mathbb{P}}$ corresponding truth assignment T^*. By Claim A.1 and A.2, it is clear that $L - q \le F(s, P^*\hat{b}) \le L$. Let x^* be the matching solution of b-matching problem $F(s, P^*\hat{b})$. For each clause i, we have $\sum_{e \in \{\{v_i^c, v_j^u\}, \{v_i^c, v_j^{\neg u}\} \mid j = 1,\ldots,p\}} x_e^* = 1$, because every clause in C is true and $(P^*\hat{b})_{v_i^c} = q + 1$. By Claim A.3, this implies $F(s, P^*\hat{b}) = L$.

For a necessary condition, assume that C is not satisfiable. We should show that for any $P \in \mathbb{P}$ we have $F(s, P\hat{b}) < L$. Assume that, for contradiction, there exists $\tilde{P}$ such that $F(s, \tilde{P}\hat{b}) = L$. By Claim A.2, $\tilde{P} \in \hat{\mathbb{P}}$, and by Claim A.3, we have $\sum_{e \in E_{c,u} \cup E_{c,\neg u}} \tilde{x}_e = q$ where $\tilde{x}$ is a solution of problem $F(s, \tilde{P}\hat{b})$. This means we have a truth assignment $T_{\tilde{P}}$ that satisfies every clause in C which derives a contradiction. □

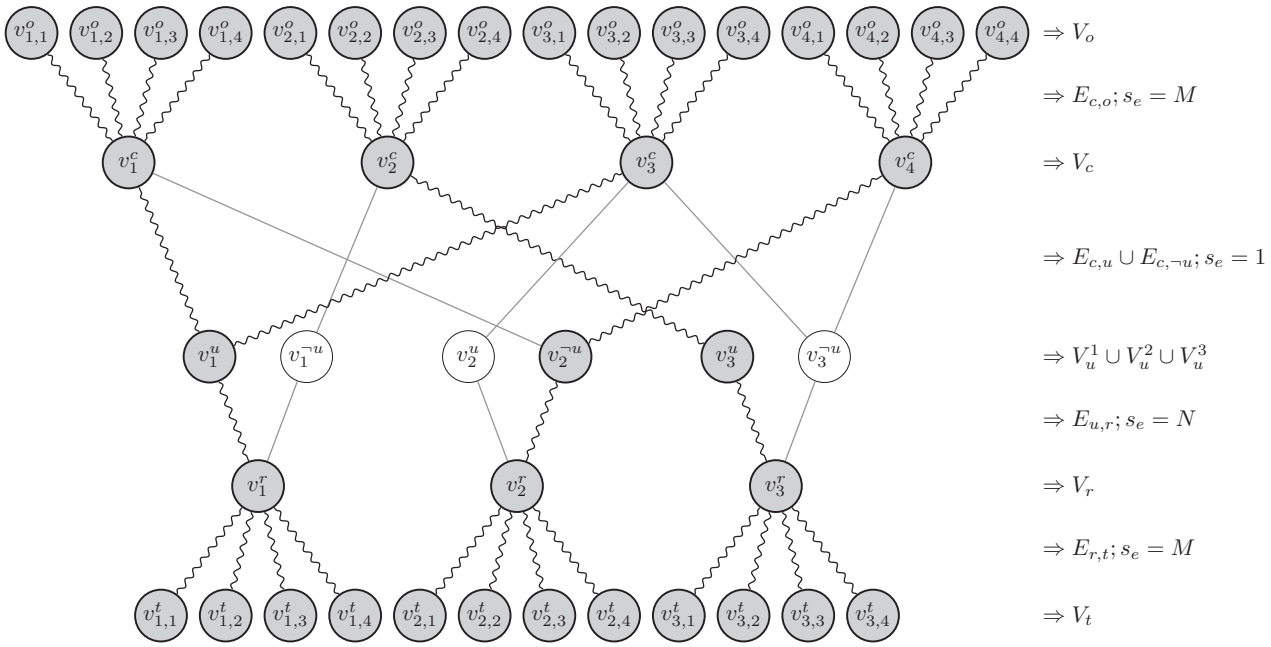

Figure A.1 **Reduction from SAT instance** $(u_1 \vee \neg u_2) \wedge (\neg u_1 \vee u_3) \wedge (u_1 \vee u_2 \vee \neg u_3) \wedge (\neg u_2 \vee \neg u_3)$**. Dark nodes (⊙) have node degree constraints** $q+1=5$ **and bright nodes (○) have node degree constraints 0. The wavy edges are the maximum weight** b**-matching for the current permutation of node degree constraints. The truth assignments for the SAT instance are** $u_1 = true$**,** $u_2 = false$**, and** $u_3 = true$**.**

Appendix B: Additional AUC Results for Tested Networks

Table B.1 – B.4 show monthly AUC results for testes networks. We included the p-values of the paired and one-sided student t-test with the alternative hypothesis: the average AUC of DD approach is better than the average of SO method. The lower p-value the more likely the DD approach performed better.

References

Anstee, R. P. 1987. A polynomial algorithm for b-matchings: an alternative approach. *Information Processing Letters* **24** 153–157.

Table B.1 Link prediction results (AUC values) for the Enron e-mail networks.

month-year	$\|E_t\|$	$\hat{\alpha}$	Static		Time-series		HRM		All	
			SO_{ST}	DD_{ST}	SO_{TS}	DD_{TS}	SO_{HRM}	DD_{HRM}	SO_{ALL}	DD_{ALL}
5-2000	46	1.51	**0.9585**	0.9486	0.9256	0.9177	0.9389	0.9194	0.9581	0.9498
6-2000	66	1.51	**0.8698**	0.8544	0.8415	0.8342	0.8686	0.8553	0.8686	0.8547
7-2000	82	1.51	0.8826	0.9073	0.7720	0.8048	0.8880	**0.9080**	0.8870	0.9058
8-2000	129	1.42	0.8844	0.8783	0.7813	0.7517	0.8784	0.8667	**0.8845**	0.8766
9-2000	100	1.40	0.9648	0.9522	0.8939	0.8545	0.9509	0.9420	**0.9726**	0.9610
10-2000	141	1.38	0.8804	0.9019	0.8238	0.8543	0.8723	0.8798	0.8846	**0.9026**
11-2000	165	1.34	0.9507	0.9511	0.8431	0.8537	0.9253	0.9133	**0.9523**	0.9509
12-2000	164	1.45	0.9455	0.9473	0.8611	0.9036	0.9283	0.9184	**0.9520**	0.9493
1-2001	148	1.61	0.9258	0.9235	0.8339	0.8392	0.8558	0.8478	**0.9269**	0.9248
2-2001	172	1.58	0.9606	0.9552	0.8836	0.8993	0.9472	0.9336	**0.9647**	0.9568
3-2001	184	1.50	0.9700	0.9644	0.9017	0.9021	0.9389	0.9269	**0.9745**	0.9670
4-2001	212	1.44	0.9039	0.9127	0.8151	0.8329	0.8806	0.8809	0.9072	**0.9146**
5-2001	249	1.36	0.8163	**0.8252**	0.7459	0.7648	0.8018	0.7984	0.8210	0.8250
6-2001	197	1.34	0.8146	**0.8214**	0.7722	0.7649	0.7736	0.7730	0.8123	0.8146
7-2001	219	1.41	0.8381	0.8529	0.8067	0.7911	0.8679	0.8625	0.8635	**0.8712**
8-2001	342	1.38	0.8534	0.8591	0.7591	0.7657	0.8146	0.8192	0.8634	**0.8668**
9-2001	303	1.29	0.8800	0.9137	0.9035	0.8830	0.8631	0.8587	0.9118	**0.9192**
10-2001	490	1.33	0.8679	0.8950	0.8036	0.8011	0.8327	0.8419	0.8939	**0.8977**
11-2001	410	1.13	0.8358	0.8861	0.8572	0.8570	0.8786	0.8727	0.9060	**0.9148**
12-2001	279	1.04	0.8594	0.9062	0.9085	0.9003	0.8916	0.8857	0.9318	**0.9339**
1-2002	275	0.99	0.8295	0.8850	0.8403	0.8386	0.8457	0.8427	0.9077	**0.9156**
2-2002	246	1.09	0.8245	0.8511	0.7652	0.8313	0.7890	0.7844	0.8727	**0.8739**
3-2002	72	1.29	0.8298	0.9028	0.9299	0.9281	0.8441	0.8462	0.9313	**0.9395**
average	204.0	1.36	0.8846	0.8998	0.8378	0.8423	0.8729	0.8686	0.9065	**0.9081**
p-value				0.0030		0.1804		0.9851		0.1865

Table B.2 Link prediction results (AUC values) for the stock correlation networks.

month-year	$\lvert E_t\rvert$	$\hat{\alpha}$	Static		Time-series		HRM		All	
			SO_{ST}	DD_{ST}	SO_{TS}	DD_{TS}	SO_{HRM}	DD_{HRM}	SO_{ALL}	DD_{ALL}
1-2009	3837	1.22	0.9111	0.8976	0.8417	0.8362	0.8895	0.8846	**0.9185**	0.9035
2-2009	1349	1.24	0.8502	0.8626	0.8424	0.8367	0.8742	0.8665	**0.8809**	0.8747
3-2009	1962	1.22	0.8853	0.8905	0.8647	0.8600	0.9005	0.8872	**0.9163**	0.9061
4-2009	3382	1.20	0.9219	0.9099	0.8506	0.8439	0.9095	0.8905	**0.9364**	0.9180
5-2009	2809	1.16	0.8828	0.8822	0.8671	0.8594	0.9051	0.8919	**0.9129**	0.9010
6-2009	1032	1.12	0.8803	0.9005	0.8927	0.8871	0.9175	0.8957	**0.9277**	0.9224
7-2009	701	1.12	0.8519	0.8620	0.8699	0.8646	0.8899	0.8689	**0.8907**	0.8853
8-2009	1229	1.15	0.8881	0.8952	0.8879	0.8815	0.8924	0.8697	**0.9136**	0.9049
9-2009	401	1.15	0.8128	0.8431	0.8606	0.8618	0.8785	0.8662	0.8772	**0.8796**
10-2009	873	1.23	0.8890	0.8974	0.8799	0.8727	0.9207	0.9022	**0.9219**	0.9140
11-2009	609	1.25	0.8140	0.8387	0.8402	0.8380	**0.8855**	0.8721	0.8718	0.8700
12-2009	382	1.30	0.8281	0.8568	0.8771	0.8763	**0.8861**	0.8782	0.8807	0.8853
1-2010	458	1.32	0.7625	0.7789	0.7837	0.7833	**0.8142**	0.8028	0.8098	0.8127
2-2010	432	1.37	0.7084	0.7427	0.7320	0.7321	0.7725	0.7667	0.7721	**0.7756**
3-2010	173	1.44	0.7866	0.8620	0.8426	0.8460	0.8857	0.8909	0.8841	**0.8958**
4-2010	835	1.56	0.8505	0.8557	0.8081	0.8059	**0.8851**	0.8740	0.8823	0.8786
5-2010	2031	1.59	0.8770	0.8761	0.8245	0.8218	0.8856	0.8779	**0.9001**	0.8913
6-2010	1390	1.65	0.9206	0.9175	0.8528	0.8508	0.9229	0.9124	**0.9397**	0.9303
7-2010	979	1.65	0.8272	0.8337	0.8069	0.8077	**0.8604**	0.8599	0.8566	0.8550
8-2010	387	1.64	0.8261	0.8320	0.8040	0.8047	0.8542	0.8540	**0.8568**	0.8533
9-2010	577	1.65	0.8726	0.8796	0.8165	0.8164	0.8757	0.8743	**0.8882**	0.8854
10-2010	349	1.60	0.6839	0.7006	0.7173	0.7209	**0.7746**	0.7616	0.7422	0.7457
11-2010	499	1.63	0.8399	0.8458	0.8037	0.8045	**0.8822**	0.8700	0.8815	0.8776
12-2010	247	1.61	0.7777	0.8106	0.7754	0.7742	0.8195	0.8204	0.8293	**0.8350**
average	1121.8	1.38	0.8395	0.8530	0.8309	0.8286	0.8743	0.8641	**0.8788**	0.8751
$\lvert E_t\rvert < 500$			0.7807	0.8081	0.7996	0.8004	0.8408	0.8345	0.8371	0.8401
p-value				0.0023		0.1011		0.9883		0.0447
$500 \le \lvert E_t\rvert < 1000$			0.8509	0.8612	0.8369	0.8342	0.8862	0.8753	0.8852	0.8814
p-value				0.0087		0.0417		0.9874		0.9944
$1000 \le \lvert E_t\rvert < 1500$			0.8848	0.8940	0.8689	0.8640	0.9017	0.8861	0.9155	0.9081
p-value				0.0786		0.9919		0.9867		0.9977
$1500 \le \lvert E_t\rvert$			0.8956	0.8913	0.8498	0.8442	0.8981	0.8864	0.9169	0.9040
p-value				0.8511		0.9985		0.9955		0.9991

Table B.3 Link prediction results (AUC values) for the Facebook500 friend network.

month-year	new $\lvert E_t \rvert$	$\hat{\alpha}$	Static		HRM		All	
			SO_{ST}	DD_{ST}	SO_{HRM}	DD_{HRM}	SO_{ALL}	DD_{ALL}
9-2006∼10-2006	54	0.98	0.9336	0.9529	**0.9751**	0.9727	0.9638	0.9721
11-2006∼12-2006	27	1.18	0.9030	0.9268	0.8602	0.8607	0.9161	**0.9344**
1-2007∼2-2007	23	1.31	0.9118	**0.9314**	0.9161	0.9122	0.9186	0.9255
3-2007∼4-2007	26	1.38	0.9111	0.9392	0.9450	**0.9525**	0.9216	0.9432
5-2007∼6-2007	28	1.46	0.9139	0.9406	0.9202	0.9232	0.9378	**0.9435**
7-2007∼8-2007	28	1.53	0.9189	0.9471	0.8887	0.8867	0.9525	**0.9652**
9-2007∼10-2007	41	1.59	0.8335	0.8733	0.8285	0.8222	0.8574	**0.8762**
11-2007∼12-2007	44	1.64	0.8781	0.9023	0.8670	0.8696	0.9115	**0.9143**
1-2008∼2-2008	26	1.70	0.7744	0.7934	0.7220	0.6993	0.7978	**0.8026**
3-2008∼4-2008	77	1.74	0.8761	0.8963	0.8803	0.8789	0.9122	**0.9165**
5-2008∼6-2008	45	1.81	0.8433	0.8594	**0.8799**	0.8756	0.8620	0.8667
7-2008∼8-2008	43	1.86	0.8281	0.9138	0.8903	0.8984	0.8832	**0.9241**
9-2008∼10-2008	51	1.90	0.8249	0.8781	0.8312	0.8338	0.8586	**0.8784**
11-2008∼12-2008	47	1.93	0.8425	**0.8888**	0.8148	0.8160	0.8712	0.8832
average	40.0	1.57	0.8709	0.9031	0.8728	0.8716	0.8975	**0.9104**
p-value				0.0000		0.7291		0.0002

Table B.4 Link prediction results (AUC values) for the Facebook1000 friend network.

month-year	new $\lvert E_t \rvert$	$\hat{\alpha}$	Static		HRM		All	
			SO_{ST}	DD_{ST}	SO_{HRM}	DD_{HRM}	SO_{ALL}	DD_{ALL}
9-2006∼10-2006	110	0.63	0.9205	**0.9421**	0.8481	0.8461	0.9169	0.9338
11-2006∼12-2006	61	0.86	0.8763	**0.9027**	0.8435	0.8432	0.8752	0.8980
1-2007∼2-2007	63	1.01	0.9265	**0.9478**	0.8827	0.8825	0.9184	0.9375
3-2007∼4-2007	74	1.10	0.8869	**0.9082**	0.7120	0.7026	0.8789	0.9049
5-2007∼6-2007	75	1.20	0.9132	**0.9373**	0.7992	0.8005	0.9103	0.9315
7-2007∼8-2007	123	1.27	0.8977	**0.9222**	0.8152	0.8140	0.8915	0.9106
9-2007∼10-2007	108	1.35	0.8446	**0.8859**	0.7572	0.7529	0.8446	0.8805
11-2007∼12-2007	97	1.41	0.8419	**0.8653**	0.7868	0.7761	0.8419	0.8616
1-2008∼2-2008	104	1.46	0.7476	**0.8180**	0.7310	0.7190	0.7476	0.7815
3-2008∼4-2008	207	1.52	0.8898	0.9223	0.7399	0.7405	0.8898	**0.9230**
5-2008∼6-2008	143	1.58	0.8341	**0.8924**	0.7799	0.7745	0.8341	0.8803
7-2008∼8-2008	130	1.64	0.7998	0.8671	0.7074	0.7111	0.7998	**0.8675**
9-2008∼10-2008	138	1.69	0.8228	**0.8740**	0.7138	0.7162	0.8228	0.8728
11-2008∼12-2008	167	1.73	0.7587	0.8131	0.6591	0.6616	0.7587	**0.8135**
average	114.3	1.32	0.8543	**0.8927**	0.7697	0.7672	0.8522	0.8855
p-value				0.0000		0.9534		0.0000