DOI 10.26886/2311-4517.4(89)2023.3

УДК 004.724.2

ВИЯВЛЕННЯ ВІРУСНИХ «ХРОБАКІВ» ЕЛЕКТРОННОЇ ПОШТИ ЗА ДОПОМОГОЮ ВЕЙВЛЕТ-АНАЛІЗУ ПОТОКІВ ЗАПИТІВ DNS.

Гребеннік Кирил Валентинович, магістрант

http://orcid.org/0009-0002-3958-3912

e- mail: kyryl.hrebennik@nure.ua

Харківський національний університет радіоелектроніки, Україна, Харків

Представлено дослідження виявлення поштових черв'яків за допомогою вейвлет-перетворень. Стаття робить крок до кращого розуміння черв'яків електронної пошти та дослідження їх впливу на рівень характеристик потоків запитів системи доменних імен (DNS), які створюють машини користувачів. Для моделювання та експериментальних обчислень використано вейвлет-аналіз, а саме дискретне та неперевне вейвлет перетворення, статистичні алгоритми кластеризації, численні методи та інші методи математичного аналізу.

Ключові слова: поштові черв'яки; DNS запити; дискретне вейвлетне перетворення; вейвлет Хаара (Гаара), стиснення даних.

магистрант Гребенник К.В. выявление вирусных «червяков» электронной почты с помощью вейвлет-анализа потоков запросов DNS / Харьковский национальный университет радиоэлектроники, Харьков, Украина.

Представлены исследования обнаружения почтовых червей с помощью вейвлет-преобразований. Статья делает шаг к лучшему пониманию червей электронной почты и исследованию их влияния на уровень характеристик потоков запросов системы доменных

имен (DNS), создающимся пользовательскими машинами. Для моделирования и экспериментальных вычислений использованы вейвлет-анализ, а именно дискретное и непрерывное вейвлет преобразование, статистические алгоритмы кластеризации, численные методы и другие методы математического анализа.

Ключевые слова: почтовые черви; DNS запросы; дискретное вейвлетное превращение; вейвлет Хаара, сжатие даннях undergraduate Grebennik K.V. detection of viral email worms using wavelet analysis of DNS query streams / Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Investigations into detection of mail worms using wavelet transforms are presented. The article takes a step towards a better understanding of email worms and the study of their impact on the level of performance of Domain Name System (DNS) query flows generated by user machines. Wavelet analysis, namely discrete and continuous wavelet transform, statistical clustering algorithms, numerical methods and other methods of mathematical analysis, were used for modeling and experimental calculations.

Key words: mail worms; DNS requests; discrete wavelet transformation; Haar wavelet, data compression

Вступ. Інтернет-хробаки сьогодні є одними з основних оперативних викликів безпеки, зі спалахами зараженнями черв’яків пов’язані величезні грошові втрати. Надалі терміном «хробак» (черв’як) використовуватиметься, щоб охопити кожну шкідливу програму, яка поширюється через комп'ютерну мережу незалежно від взаємодії людини потрібен (вірус) чи ні (хробак). На основі їх розмноження, хробаки поділяються на скануючі (на основі експлойтів) і топологічні [1 с. 11–18; 2; 3; 4; 5 с. 187–221]. Черв'яки-сканери

використовують вразливість, щоб заразити машину користувача, а потім поширити за адресами вибрати з простору IP-адрес. Топологічні черви покладаються головним чином на соціальній інженерії для зараження комп'ютера користувача та використовувати інформацію, яку вони збирають з машини, для поширення серед соціальних контактів. Соціальна інженерія – це вид вторгнення, що сильно залежить від взаємодії людини, що передбачає переслідування користувачів з метою порушення нормальної безпеки.

Ресурсні записи DNS – записи про відповідність імені і службової інформації в системі доменних імен. Оскільки користувачі стають все більш обізнаними про загрозу електронної пошти складність механізмів захисту хробаків збільшується, автори хробаків стають все більш занепокоєними про збільшення рівня їх зараження. Тому, для того, щоб швидкість хробаків досягнула рівня епідемії, вони їх споряджають агресивними методами збирання списків з багатьма адресами електронної пошти. Тобто, підтверджується гіпотеза про те, що потоки запитів DNS діляться на два канонічні профілі. користувачів і інший – на машини, заражені різними поштовими хробаками.

Підготовка даних та пошук подібності часових рядів. Як вхідні дані метод використовує набір DNS-запитів такі, як локальний сервер імен, отриманий протягом інтервалу спостереження T. Кожен запит складається з часу, коли сервер отримав запиту, IP-адресу хоста, що запитує, і запитуваного даних. Запити групуються за запитуючим хостом, і оскільки для дослідження не потрібна інформація прикладного рівня, зберігаються лише час запиту та IP-адреса запитувача хост. Для кожного хоста розглядається p послідовних інтервалів часу однакової ширини, і для кожного ящика підраховуються запити. У цьому процесі отримується набір

однофакторних часових рядів, де кожен із них виражає кількість DNS-запитів на хост через певний час. Набір часових рядів можна виразити як матриця часових рядів n × p; n – кількість хостів, які запитав сервер імен принаймні один раз у межах T.

Кожен часовий ряд із p часовими точками можна розглядати як точку у p-вимірному просторі. Це дозволяє використовувати багатоваріантність алгоритмів аналізу даних безпосередньо для дослідження даних часових рядів. Однак більшість алгоритмів інтелектуального аналізу даних, зокрема, більшість класичних алгоритмів кластеризації не працюють належним чином з рядами. Робота з кожною точкою часу робить значення метрик відстані, які використовуються для вимірювання схожість між об'єктами сумнівна [6 с. 420–434]. Це означає що пошук значущих груп є значно складнішим і неочевидним, оскільки кластеризація втрачає свою алгоритмічну ефективність із збільшенням розмірності даних. До вирішення цієї проблеми, існують численні представлення часових рядів, які полегшують вилучення стисненого низькорозмірного представлення часового ряду (тобто вектор ознак), який служить вхідними даними для алгоритмів аналізу даних.

На вибір репрезентації сильно впливає завдання аналізу даних і мета подібності. Два види подібності обговорюються в літературі часових рядів: подібність за формою та структурна подібність. На основі форми подібність використовує вихідні дані для пошуку гомоморфних послідовностей, тоді як структурна подібність описує її в автокореляційній структурі [7 с. 11–40]. У цій статті розглянуто саме подібність форми, щоб знайти подібні потоки запитів. Хоча було запропоновано багато уявлень, порівнюючи часові ряди за формою, лише деякі з них підходять. для даної вибірки. Кластеризація векторів ознак це не оригінальний часовий ряд, який потенційно може ввести

фальшиві відхилення. Помилкові відхилення відбуваються, коли вектори ознак подібних часових рядів віддалені у векторному просторі ознак. У даній структурі використано представлення, яке гарантує відсутність фальшивих відхилень.

Фалуцос та ін. [8 с. 419–429] представив GEneric Multimedia Method INdexIng (GEMINI), який визначає умови представлення, щоб відповідати подібності пошуку, який не передбачає помилкових відхиленнь (похибок). GEMINI – це трикрокова процедура. По-перше, метрика відстані *dts* у часі необхідно визначити простір серії. Потім дані часових рядів трансформуються та стискаються, щоб отримати вектори ознак. Робота з векторами ознак, а не з оригінальним часом серій гарантує відсутність помилкових відхилень, якщо можна визначити міру відстані dfv на векторах ознак, що відповідає:

$$d_{fv}\left(f(ts_i), f\left(ts_j\right)\right) \leq d_{ts}\left(ts_i, ts_j\right) \quad (1)$$

tsi, *tsj* – часові ряди, а *f()* – низьковимірне представлення функції вилучення. Ця лема відома як лема про нижню межу. Оригінальна робота в [8] використовує Дискретне перетворення Фур'є (ДПФ) вихідного часового ряду, але подальша робота показала, що інші представленя задовольняють рівняння 1. Джерело [7] дає повний огляд часових рядів, виділяючи малу підмножину дійснозначних представлень, які задовольняє рівняння. 1.

Вейвлет-перетворення та стиснення даних. Для даного аналізу використовується дискретне вейвлетне перетворення (DWT), яке апроксимує часовий ряд суперпозицією базису функції. Ці базисні функції утворюються шляхом розширення і трансляції базисної вейвлет-функції. DWT представлення за своєю суттю має багатороздільну здатність і дозволяє одночасний аналіз часу та

частоти, оскільки він перетворює часові ряди на коефіцієнти, локалізовані в часі. Це дозволяє відстежувати зміни в характеристиках часу рядів в певному масштабі як функція часу. Крім того, для часових рядів, які зазвичай зустрічаються на практиці, багато з коефіцієнтів дорівнюють нулю або дуже малі, що дозволяє ефективне стиснення. Крім загальних переваг DWT, ще два фактори мотивують його використання. По-перше, DWT застосована для аналізу нестаціонарних сигналів, тобто сигналів, частотний зміст яких змінюється в часі; і по-друге, це добре працює при стисненні розріджених часових рядів. DWT застосовується незалежно до кожного часового ряду матриця часових рядів з використанням алгоритму піраміди Маллата – базовий метод, доступний у будь-якому статистичному програмному забезпеченні. Алгоритм Маллата розкладає p довжину – з p є ступенем двох часових рядів до p вейвлет-коефіцієнтів у розкладі $log_2 p$ рівнів, де кожен рівень відповідає діапазону частот. Тому після нанесення DWT на кожну лінію часу матриці серій, отримується матриця вейвлет-коефіцієнтів $n \times p$.

Щоб зменшити розмірність вейвлет-коефіцієнтів матриці, застосовується техніка стиснення, яка вибирає невелика підмножина вейвлет-коефіцієнтів, які забезпечують високу дискримінаційну силу між часовими рядами та хорошою кластеризацією. Для цілей стиснення вейвлет-коефіцієнти часто нормалізуються, що означає, що коефіцієнти при нижчій роздільній здатності мають більшу вагу, ніж коефіцієнти при вищій роздільній здатності. Зберігаючи k найбільших коефіцієнтів у терміни абсолютного нормалізованого значення, дає для даного бюджету коефіцієнтів k оптимальне вейвлет-представлення в термінах помилки суми квадратів [9, с.36–41]. Ще одна добре відпрацьована техніка пропонує зберегти перші k коефіцієнтів, які описують низькочастотні особливості часового ряду.

Mörchen [10] порівнює ці дві методики, які застосовуються незалежно для кожного часового ряду з двома методами, застосовними до набору n часових рядів. Ці прийоми застосовуються безпосередньо до матриці вейвлет-коефіцієнтів.

Перший зберігає k стовпців матриці, яка має найбільшу середню в квадраті елементів вартість; тоді, як другий для даного k є n × k найбільшим коефіцієнтом матриці вейвлет-коефіцієнтів. Збереження перших k вейвлет-коефіцієнтів або k стовпців матриці вейвлет-коефіцієнтів, які мають найбільше середнє значення поелементне квадратне значення створює вектор ознак n × k матриця, з k << p. Тоді як два інших компресійні методи створюють векторну матрицю ознак n × p із n × k ненульових елементів. На практиці використовується одна техніка стиснення [11, с. 60–63].

У математиці Гаарів вейвлет (англ. Haar wavelet) – найпростіший випадок дискретного вейвлетного перетворення, це послідовність перемасштабованих функцій «квадратної» форми, які разом утворюють вейвлетне сімейство або базис. Вейвлетний аналіз подібний до аналізу Фур'є тим, що дозволяє цільовій функції над інтервалом бути поданою в термінах ортонормованого базису. Гаарову послідовність тепер визнають як перший відомий вейвлетний базис, та широко використовують як навчальний приклад.

Таблиця 1

Середнє (μ) та стандартне відхилення (δ) по десяти наборах даних від % поштових черв'яків, для яких правила зупинки для коєфіцієнтів k=4, k=8 Connectivity (C.), Dunn (D.), Silhouette (S.) і Davies-Bouldin(DB) вкажуть, що лише дві різних основні вибірки існують у векторній матриці характеристик для чотирьох технік стискання даних.

		k=4				k=8			
		C.	D.	S.	DB	C.	D.	S.	DB
FC	μ	100	93.3	97.5	92.4	100	94.8	99.5	94.2
	σ	0	13.6	3	13.4	0	11.2	1	11.1
LC	μ	100	98	98.6	95.4	100	98.8	99.1	97.5
	σ	0	2.7	2.6	10.4	0	2.1	1.9	7.5
BKC	μ	100	97.0	98	94.9	100	96.3	98.9	98.8
	σ	0	5.9	1.9	9.9	0	9.7	1.8	11.0
BCO	μ	100	97.9	98.9	97.6	100	98	99.2	97.7
	σ	0	3.8	2.3	3.8	0	3.7	1.7	3.7

Таблиця 2

Середнє (μ) та стандартне відхилення (δ) по десяти наборах даних від % поштових черв'яків, для яких правила зупинки для коєфіцієнтів k=16, k=32 Connectivity (C.), Dunn (D.), Silhouette (S.) і Davies-Bouldin(DB) вкажуть, що лише дві різних основні вибірки існують у векторній матриці характеристик для чотирьох технік стискання даних. усі правила розкривають явне домінування двокластерної схеми.

		k=16				k=32			
		C.	D.	S.	DB	C.	D.	S.	DB
FC	μ	100	98.7	99.7	95.6	100	96.5	99.9	96.2
	σ	0	1.7	0.6	9.9	0	9.7	0.4	9.7
LC	μ	100	99.1	99.2	97.3	100	99	99.1	97.4
	σ	0	1.8	1.7	7.5	0	1.9	1.8	6.7
BKC	μ	100	96.1	97.9	95.6	100	96.6	98.3	95.3
	σ	0	9.6	2.8	9.6	0	8.3	1.9	8.3
BCO	μ	100	98	99.2	97.8	100	98.1	99.3	97.9
	σ	0	3.7	1.5	3.7	0	3.5	1.3	3.6

Ієрархічна кластеризація та правила зупинки. У даній роботі кластеризуються рядки векторної матриці ознак, щоб зробити висновок як про кількість, так і про природу окремих вибірок. Число вказує на кількість окремих DNS класів активності, які існують у досліджених потоках запитів DNS. Мета може бути математично виражена як доказ того, що найкраща кластеризація ділить вектори ознак на два кластери. Щодо природи цих класів активності DNS, враховуючи двокластерну схему, показується, що один кластер містить лише вектори функцій неінфікованих машин користувача та

інші вектори функцій електронної пошти, заражені хробаками.

Експериментальна оцінка та результати обчислення. У таблицях 1, 2 показується правила зупинки значення k, техніка стиснення та середнє значення (над десятьма наборами даних) і стандартне відхилення відсотка черв'яків електронної пошти для яких правила зупинки вказують, що двокластерна схема є найкращою кластеризацією. У таблиці FC, LC, BKC, і BCO стосується збереження перших k коефіцієнтів, найбільшого k, які мають найбільші середньоквадратичні значення та найбільші n×k коефіцієнти матриці вейвлет-коефіцієнтів відповідно.

Для, наприклад, значення 100 у першій верхній лівій клітинці таблиці дорівнює інтерпретувати таким чином: зв'язок показує, що в середньому для наборів даних двокластерна схема є найкращою кластеризацією для всіх – тобто 100% – перевірених черв'яків електронної пошти, коли перші чотири коефіцієнти на часовий ряд зберігаються. У таблиці показується, що всі чотири правила зупинки демонструють явне домінування двокластерної схеми над будь-яким іншим результатом кластеризації. Це означає, що індекси розкривають лише два відмінних основні сукупності існують у векторній матриці ознак. Зокрема, збереження найбільших n×k коефіцієнтів перевершує інші методи стиснення, оскільки для кожного k усі правила вказують на те, що в середньому понад 97% email черв'яків існують точно два канонічні профілі [12, с. 28–33]. Тому надалі, надається перевага виключно цій техніці стиснення.

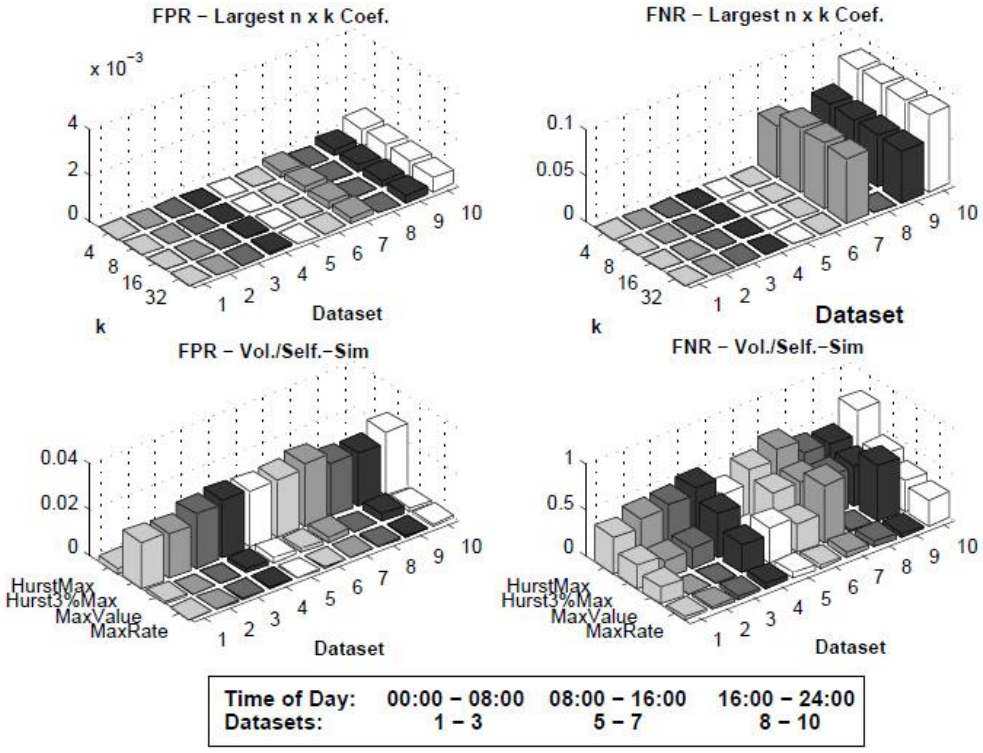

Рисунок 1 – Частота хибнопозитивних (FPR) і хибнонегативних (FNR) вибірок різних поштових хробаків з різними наборами даних (верхні графіки) і порівняння даного методу із методами, які висвітлюють виключно по обсягу або самоподібності DNS трафіку (нижні графіки).

Другий метод на основі порогового значення обсягу, позначається MaxValue, виявляє, що заражена email-хробаком машина користувача, яка генерує максимальну кількість запитів вимірюється в 15-секундному діапазоні часу. Третій спосіб натхненний результатами в [13, с. 67–72], які показують, що атака шаблоне сканування хробаків є самоподібною. Це свідчить про те, що електронна пошта, заражена хробаком, генерує DNS-трафік, який має найвищий ступінь самоподібності. Для вимірювання ступеня самоподібності оцінюється параметр H за допомогою

непараметричної оцінки [14, с.39–44].

Висновки. Таким чином, на основі вищевикладеного, знятих даних в результаті експериментів, та виконаних обчислень можна зробити висновки, що заявлений метод виявлення вірусів може бути використаний для виявлення активності електронного хробака у відправника, швидше, ніж в одержувача листів домену. Це означає, що він більш ефективніший, ніж подібні методи, засновані на самоподібності вірусу або накопиченні обсягу запитів.

Література:

1. N. Weaver, V. Paxson, S. Staniford, and R. Cunningham, “A taxonomy of computer worms,” in WORM ’03: Proc. of the 2003 ACM workshop on Rapid malcode. New York, NY, USA: ACM, 2003, pp. 11–18.

2. M. Braverman, “Behavioral modeling of social engineering-based malicious software,” in Virus Bulletin Conf., 2006.

3. Virus Radar, “Top 10 threats,” <http://www.virus-radar.com>

4. Kaspersky Lab, “Monthly Malware Statistics,” <http://www.viruslist.com>.

5. S. Stolfo, S. Hershkop, C. Hu, W. Li, O. Nimeskern, and K. Wang, “Behavior-based modeling and its application to email analysis,” ACM Trans. Interet Technol., vol. 6, no. 2, pp. 187–221, 2006.

6. C. Aggarwal, A. Hinneburg, and D. Keim, “On the Surprising Behavior
of Distance Metrics in High Dimensional Space,” in ICDT 2001: Proc. of the 8th Int. Conf. on Database Theory, ser. LNCS. Springer, 2001, pp. 420–434

7. A. Bagnall, C. Ratanamahatana, E. Keogh, S. Lonardi, and G. Janacek, “A bit level representation for time series data mining with shape based similarity,” Data Mining Knowledge Discovery, vol. 13, no. 1, pp.

11–40, 2006.

8. C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, “Fast subsequence matching in time-series databases,” in SIGMOD ’94: Proc. of the 1994 ACM SIGMOD Int. Conf. on Management of Data. ACM, 1994, pp. 419–429

9. N. Chatzis, “Motivation for behaviour-based dns security: A taxonomy of dns-related internet threats,” in SECURWARE 2007: Proc. of the Int. Conf. on Emerging Security Information, Systems, and Technologies. Los Alamitos, CA, USA: IEEE Computer Society, 2007, pp. 36–41.

10. F. Mörchen, “Time series feature extraction for data mining using dwt and dft,” Dept. of Maths and CS, Philipps-U. Marburg, Tech. Rep. No. 33, 2003.

11. Vidakovic, Brani (2010). *Statistical Modeling by Wavelets*. Wiley Series in Probability and Statistics (вид. 2). с. 60, 63.

12. K. E. Stollnitz, T. Derose, and D. Salesin, Wavelets for computer graphics: theory and applications. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996

13. R. Matsuba, Y. Musashi, and K. Sugitani, “Detection of mass mailing worm-infected ip address by analysis of syslog for dns server,” IPSJ SIG, pp. 67–72, 2004.

14. Y. Musashi and K. Rannenberg, “Detection of mass mailing worminfected pc terminals by observing dns query access,” IPSJ SIG Notes, pp. 39–44, 2004.

References:

1. N. Weaver, V. Paxson, S. Staniford, and R. Cunningham, “A taxonomy of computer worms,” in WORM ’03: Proc. of the 2003 ACM workshop on Rapid malcode. New York, NY, USA: ACM, 2003.

2. M. Braverman, “Behavioral modeling of social engineering-based

malicious software," in Virus Bulletin Conf., 2006.

3. Virus Radar, "Top 10 threats," <http://www.virus-radar.com>.

4. Kaspersky Lab, "Monthly Malware Statistics," <http://www.viruslist.com>.

5. S. Stolfo, S. Hershkop, C. Hu, W. Li, O. Nimeskern, and K. Wang, "Behavior-based modeling and its application to email analysis," ACM Trans. Interet Technol., vol. 6, no. 2 2006.

6. C. Aggarwal, A. Hinneburg, and D. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," in ICDT 2001: Proc. of the 8th Int. Conf. on Database Theory, ser. LNCS. Springer, 2001.

7. A. Bagnall, C. Ratanamahatana, E. Keogh, S. Lonardi, and G. Janacek, "A bit level representation for time series data mining with shape based similarity," Data Mining Knowledge Discovery, vol. 13, no. 1, 2006.

8. C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," in SIGMOD '94: Proc. of the 1994 ACM SIGMOD Int. Conf. on Management of Data. ACM, 1994.

9. N. Chatzis, "Motivation for behaviour-based dns security: A taxonomy of dns-related internet threats," in SECURWARE 2007: Proc. of the Int. Conf. on Emerging Security Information, Systems, and Technologies. Los Alamitos, CA, USA: IEEE Computer Society, 2007.

10. F. Mörchen, "Time series feature extraction for data mining using dwt and dft," Dept. of Maths and CS, Philipps-U. Marburg, Tech. Rep. No. 33, 2003.

11. Vidakovic, Brani (2010). *Statistical Modeling by Wavelets*. Wiley Series in Probability and Statistics (вид. 2).

12. K. E. Stollnitz, T. Derose, and D. Salesin, Wavelets for computer graphics: theory and applications. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996

13. R. Matsuba, Y. Musashi, and K. Sugitani, “Detection of mass mailing worm-infected ip address by analysis of syslog for dns server,” IPSJ SIG, 2004.

14. Y. Musashi and K. Rannenberg, “Detection of mass mailing worminfected pc terminals by observing dns query access,” IPSJ SIG Notes, 2004.