第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·博士论文· 文章编号：1000—3428(2007)18—0022—04 文献标识码：A 中图分类号：TP393.08

P2P中的文件污染与污染防治

左 敏，李建华
(上海交通大学电子工程系，上海 200240)

摘 要：借助仿真实验揭示了P2P文件分发与文件污染的规律，针对常用的污染手段提出了相应的对策。实验证明，文中提出的“提醒-删除”合作机制可以有效地增加污染的难度，降低污染的成功率，从而在一定程度上缓解了P2P网络中严峻的文件污染问题。
关键词：P2P；文件共享；内容分发；文件污染

File Pollution and Anti-pollution on P2P Networks

ZUO Min, LI Jian-hua
(Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai 200240)

【Abstract】 This paper analyses the problem of file pollution on P2P file sharing networks, presents various kinds of pollution techniques, and puts forward possible counter-measures to them. In particular, a novel cooperative “warning-delete” approach is proposed to alleviate the most serious version pollution. The effectiveness of this approach is proved by simulation experiments.
【Key words】 P2P; file sharing; content distribution; file pollution

在现有P2P网络所存在的诸多问题之中，“文件污染”问题已经成为最受关注的问题之一[1~5]。一方面，“文件污染”作为一种版权保护的辅助手段(虽然是比较消极的)而被某些版权组织所采用。从2002年起，一些公司便开始雇佣P2P污染者，在各P2P网上布满其试图保护的音乐、电影、软件的假冒伪劣版本，欲使这些网络瘫痪。例如当时最著名的专业P2P污染公司Overpeer，就曾经于2002年成功地在FastTrack上使其发布的污染版本占据了该网络中所有共享文件的一半。另一方面，“文件污染”却损害了P2P一般用户的利益，并且带来进一步的安全隐患。例如，如果恶意的用户借助文件污染技术向网络中注入伪装过的携带病毒、后门等恶意程序的文件，在目前这种缺乏管理和防御机制的P2P文件共享网络中，一般用户将很难防备；再加上P2P网络在内容分发上远比传统的集中式网络高效，这些恶意文件将得以更迅速和广泛地传播，从而造成更大的危害。

1 问题描述

为了描述问题的方便，这里首先给出一些定义：

主题集 T：每个文件主题由一个或多个搜索关键词来标识，例如由歌曲名和演唱者姓名共同标识的一首歌，就是一个“主题”。主题集T是所有文件主题的集合。

版本集V：每个版本对应一个特定的文件，它有一个唯一对应的版本标识符，该标识符一般是对整个文件进行哈希计算得到的特征码。一个主题T∈T可以有多个版本，它们共同构成该主题的版本集，记作$V^T(V^T \subseteq V)$。例如同一首歌，可以有不同位速的多个MP3版本等。

副本集 C：每个副本对应一个存储在物理介质上的文件拷贝。一个版本v∈V可以有多个副本，它们可能存储在不同的节点上。同一个版本的所有副本内容一致，并具有相同的版本标识符。

当用户搜索一个主题的时候，P2P网络中的索引服务(集中式或分散式)会将该主题对应版本集(或它的一个子集)及其中每一个版本对应的副本集(或它们的子集)返回给用户；客户端程序将版本信息显示给用户之后，由用户选定一个(或多个)版本进行下载；选定版本的副本集中包含了存储这些副本的节点的地址信息，用户可以选择其中的一个或多个副本进行下载。

文件污染一般是针对某些选定的主题进行的。文件污染者有如下3种污染方式可以选择：

(1)版本污染

污染者首先针对一个(或同时针对多个)版权受保护的内容主题制造出大量的“假冒伪劣版本”(污染版本)。例如，可以在MP3文件中插入噪声，截短或插入错误的元数据等，甚至可以插入不可解码的随机字节使其变得完全不可用。然后污染者将这些污染版本的索引信息注入目标P2P网络，并且对每个污染版本，污染者都在其污染服务器上提供大量可供下载的副本，以便使这些污染版本得到迅速和广泛的传播。

由于没有有效的识别措施和管理机制，网络中的用户在搜索相关主题时很容易被这些具有大量可下载副本的污染版本所吸引。一旦下载了污染版本而又没有及时加以检验，一般用户很可能将该版本的本地副本设置为共享，并提供给其他用户下载。如此一来，污染版本将在网络中广泛地传播开来，甚至会超过了正确版本的副本数量，最终将正确副本淹没在污染副本中，使得该主题资源变得不可用。

(2)副本污染

所谓副本污染，指的是污染者声称自己存有某个正确版本的副本，但实际传给下载者的却是错误的数据。如果这种污染者足够多，那么，即使下载者能够挑选出正确的版本，一旦误选这些污染者作为下载源，也会浪费大量的时间精力

基金项目：国家“863”计划基金资助项目(2003AA142160)
作者简介：左 敏(1978－)，女，博士研究生，主研方向：分布式系统，网络与信息安全；李建华，教授、博士生导师
收稿日期：2006-09-26 **E-mail：**zuomin@sjtu.edu.cn

和网络带宽。

(3)索引污染

最简单的文件污染方式是“索引污染”[4]，即在P2P网络的索引服务系统中注入大量虚假的记录，这些记录指向不存在的版本和/或副本。当用户按照这些记录的指示尝试下载时，将得到“无法连接”的提示。如果注入的虚假索引记录足够多，那么没有耐心的用户可能在几次失败的尝试之后放弃下载的努力。索引污染既可以针对版本也可以针对副本。它与普通的版本污染和副本污染的不同之处在于，污染者注入网络中的索引记录指向并不存在的对象，因此污染者并不需要拥有强大的污染服务器来提供大量的上传服务。

以上 3 种污染方式可能单独使用，也可能结合起来形成更为复杂和隐蔽的复合式污染方式。它们的共同目标是让 P2P 网络中充斥大量不可用的索引信息、文件版本或文件副本，从而使该网络中的资源可用性大大降低，造成它的用户群逐渐失去兴趣并最终放弃使用该网络。

2 文件污染的防治对策

对污染者而言，不同的文件污染方式执行起来的难度有所不同，所能达到的污染效果也有差别。相应地，对 P2P 网络而言，针对不同的文件污染方式，可以有不同的防治对策，这些防治措施的执行难度有所不同，所能达到的污染防治效果也有差别。以下分别给以说明。

2.1 版本污染的防治

版本污染是目前最常用的污染方式，其隐蔽性好，不容易被察觉，除污染者自己提供的上传服务之外，还可以借助正常用户的共享行为在短时间内迅速传播。

由于在污染进行之初，需要在短时间内上传这些污染版本的大量副本以达到污染的目的，因此目前的污染者通常都是拥有大中型网络服务器，能够同时提供大量上传服务的专业公司。例如 Overpeer，往往可以在 Kazaa 网络中同时模拟数千个节点来分发污染版本。

文献[5]中的研究结果表明，这种污染策略至少对 Kazaa/FastTrack 这样的二层分布式 P2P 共享网络而言，是非常有效的。正因为这种污染的存在，曾经最受欢迎的 Kazaa 网络从巅峰时(2003 年～2004 年)的 400 万同时在线用户，降到了目前的约 200 万同时在线用户，大量用户转移到了尚未被严重污染的 eDonkey 和 BitTorrent 网络中。

对于版本污染，学术界目前所建议的对策是采用“版本声誉机制”[2]，即由用户对其所下载版本的真伪进行评价，然后系统对所有用户的评价进行分析整合，计算出每个版本的声誉值，并提供给后来的下载者作为选择版本的依据。

这种版本声誉机制本意是想在用户群中形成一个合作推荐系统，但它存在 3 个弱点：(1)起效慢，需要较长时间的积累才能计算出有意义的声誉值；(2)污染者可以通过不断注入新的污染版本来逃避声誉系统的制裁[1]；(3)容易受到虚假评价的攻击，污染者很容易提供大量虚假评价来颠覆一个设计不完善的版本声誉机制(例如，文献[5]表明，Kazaa的用户评价机制被证明是无效的，因为只有大约 1%的用户提供了评价，其中还有相当一部分是污染者自己给出的虚假评价)。

为此，本文提出另一种起效快而且不容易被污染者破坏的污染防治对策，称之为“提醒-删除”机制，描述如下：

(1)每个正常的下载者都记录自己的副本来源节点，一旦发现该版本是被污染的(可能需要人工参与)，就向其来源节点发出一个警告信息；

(2)被警告的节点如果确信自己的版本无误则可以预先设定为忽略警告信息(防止污染者发出虚假的警告恶意破坏正确版本的分发)；相反，如果不确信，则设定为接到警告后暂停上传该版本的副本，并提醒用户进行人工验证；

(3)如果用户验证无误则继续参与共享，提供上传服务；反之，如果发现版本的确有误，则停止上传，删除文件，并由用户决定是否尝试下载另一个版本。

“提醒-删除”机制可以借助正常节点之间的合作来提高用户的警觉度，缩短对污染版本的察觉时间，从而有效地降低污染版本的隐蔽性，达到遏制污染的效果。下一节中将说明“警觉度”对文件污染的影响，并证明“提醒-删除”机制的有效性。

2.2 副本污染的防治

由于大多数 P2P 文件共享应用都用哈希特征码来唯一标识网络中的文件版本，并使用哈希验证的机制来保证下载副本的完整性和一致性，因此，单纯的副本污染隐蔽性较差，无法借助正常用户的共享行为进行传播，只能依靠污染者自己提供的上传服务，这就大大增加了污染的难度。

单纯的副本污染对污染者而言难度比版本污染更高，而效果却不如版本污染好，因此很少被采用。但如果正确的版本经过长时间的积累已经被用户广泛认同，那么注入新的污染版本也无法达到好的污染效果，此时，污染者要降低资源的可用性就不得不采用副本污染，因此，仍然有必要研究副本污染的防治对策。

对于副本污染，学术界目前所建议的对策是采用“节点声誉机制”[4]，即下载者根据所下载副本的真伪对为其上传数据的源节点进行评价，长期大量上传错误副本的节点将被其他用户记入“黑名单”，从而使污染源被截断。

这种节点声誉机制存在着与版本声誉机制同样的弱点。并且，由于节点行为的动态特征(有些节点可能时好时坏)，节点声誉与版本声誉相比，具有更大的不准确性，也更加难以维护。关于这个问题，目前还没有很好的解决方案。

2.3 索引污染的防治

索引污染对基于 DHT(分布式哈希表)的分布式纯 P2P 网络危害比较大，因为污染者可以插入伪装的节点直接接管要污染的主题的索引职责，然后发布大量虚假的索引信息，使得该主题资源不可用。目前的对策是采用冗余索引备份，即一个主题的索引交给多个节点负责。如此一来，污染者要达到有效的污染，就必须耗费更多的资源伪装成更多的节点。例如，目前流行的Overnet DHT网络中，一个主题的索引节点甚至可以多达 1 600 个[4]。

由于索引污染并不指向实际存在的版本或副本，因此在采用了半集中式或集中式索引系统(例如 eDonkey/eMule 和 BitTorrent 的众多发布网站)的 P2P 网络中很难产生效果，因为在索引系统的统计数据和历史记录中，虚假索引的成功上传数量为 0，很容易被用户察觉和避免。再加上一些独立的网站提供的特征码验证服务(例如 ShareReator.com)，或者发布论坛管理员的手工鉴别(SuprNova.com)，可以有效地帮助用户识别和规避污染版本。

因此，目前 eDonkey/eMule 和 BitTorrent 网络中的污染现象远没有 Kazaa/FastTrack 那么严重。但是需要指出的是，这种集中式的索引系统以及特征码验证网站也有弱点，它们较容易卷入版权法纠纷，很多先驱者例如 ShareReator.com 和 SuprNova.com 都已经被判决关闭。

在以上 3 种污染方式中，版本污染是危害最大、隐蔽性最高，也是最经常被采用的一种污染方式。其危害的根源在于，由于缺乏有效的识别机制，因此污染版本可以像正确版本一样通过正常用户无意识的共享行为得到迅速的传播。本文针对这一根源提出了一种新的防治对策——“提醒-删除”机制。下一节将通过仿真实验来证明这种对策的有效性。

3 仿真试验与结果分析

3.1 基本实验设置

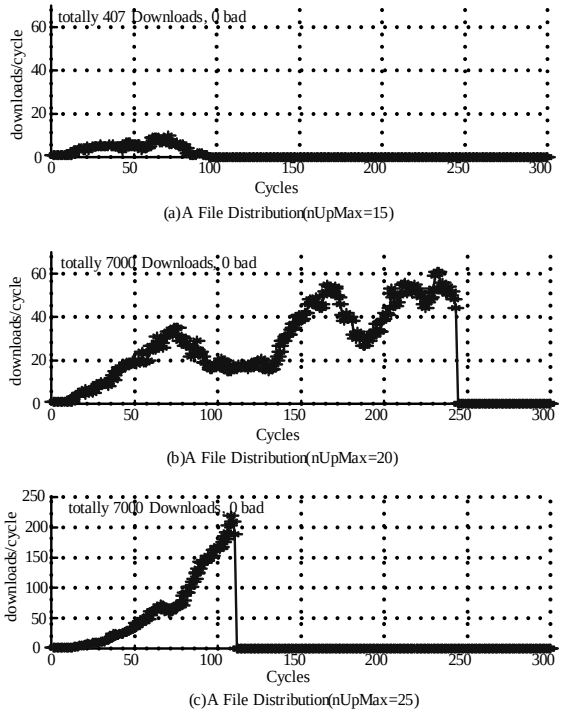
仿真实验模拟了在一个简化了的 P2P 文件共享社区中分发某个内容主题的过程。

该主题的第 1 个副本由某个随机挑选的节点(称为“发布者节点”)提供。分发过程中存在两种类型的下载者：一种是正常节点，它们完成下载后会提供一定数量的上传，每个节点的最大上传数(称之为“慷慨度”)有所不同；另一种是自私节点(许多文献中称之为“Free-riders”)，它们完成下载后立刻退出本次分发过程，不提供上传服务。

当一个节点提供上传服务的时候，称之为“种子节点”。实验中，将时间轴划分为相等的时段，称之为“轮”。每轮内，每个种子节点(含最初的发布者节点和下载成功提供上传的正常节点)可以提供一个副本的上传。假定所有的下载都是一轮一轮同步进行的，时间轴上没有交叉。

生成一个长度为 7 000 的随机队列，代表 7 000 个下载者的先后次序，这些下载者中自私节点占 90%。一般文献所报道的自私节点比例约为 70%，笔者采用了更悲观一些的配置，以增强实验结果的说服力。

图 1 显示了不同的慷慨度参数(nUpMax)对分发效率的影响，实验中正常节点p的最大上传数取[0, nUpMax]之间的一个随机整数，记为G_p。从图 1 中可以看到，节点慷慨度的高低对文件分发效率有决定性的作用：过低的慷慨度可能导致文件分发失败(图 1(a)，还有下载者尚未下载成功，就已经耗尽了种子)，慷慨度越高，每轮的种子数量越多，分发速度也越快(比较图 1(b)和图 1(c)，注意纵轴刻度的不同)。



(a)A File Distribution(nUpMax=15)

(b)A File Distribution(nUpMax=20)

(c)A File Distribution(nUpMax=25)

图 1 慷慨度对文件分发速度的影响

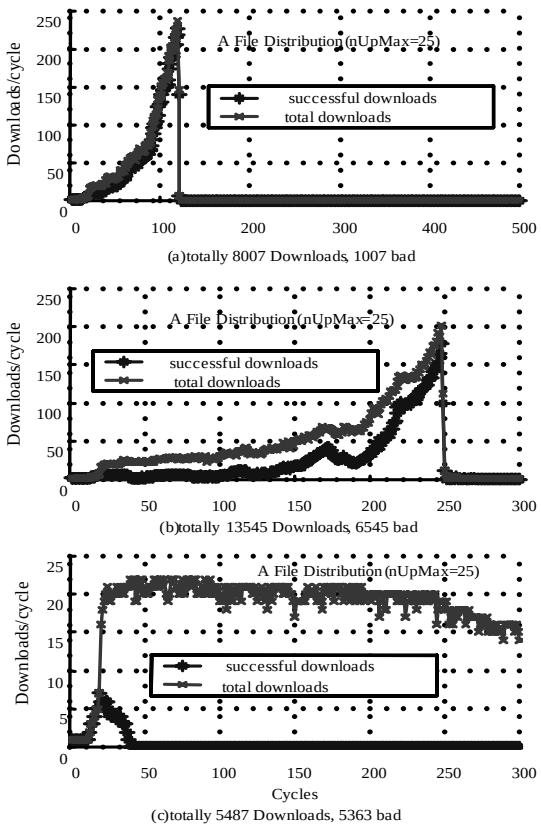
在以下的污染实验中，取 nUpMax=25 这个较高的慷慨度参数(图 1(c))。

3.2 污染实验结果

在污染实验中，假定该主题只有一个正确版本，即发布者节点最初提供的那一个，这个版本经过 20 轮的正常上传之后又获得了一些种子节点；在第 21 轮的时候，10 个污染版本的副本被同时注入网络，并开始对分发过程进行干扰。这种设置是比较保守的情况，因为实际网络中的版本污染可能远远高于 10∶1 的比率，但下面的实验可以表明，即使是这种保守的设置，文件污染的效果也非常显著。

在没有采取污染防治措施的网络中，下载者只能靠自己手工检验才能判定版本的好坏。文献[3]表明，不同的用户对版本好坏的“警觉度”是不同的，大致上是一个两极的分布。

从图 2 中可以看到警觉度参数(nAwareMax)对污染效果的影响。实验中节点的警觉度设置为：自私节点下载完成后立即检验，如果发现下载了错误的版本，则立刻将其删除并重新插入下载队列尝试下载另一个版本；一个正常节点p的警觉度参数是在[0, nAwareMax]之间随机产生的一个整数，记为A_p，该节点需要经过A_p个“轮”才会发现自己的副本是好是坏，如果是好的，则继续上传直到完成G_p个副本的传播任务，如果是坏的，则停止上传，删除副本，然后重新插入下载队列。图 2(a)nAwareMax =0，图 2(b)nAwareMax =nUpMax/2，图 2(c)nAwareMax =nUpMax。可以看出，警觉度越高(nAwareMax参数值越小意味着越高的警觉度)，分发过程中产生的额外通信量越少，污染的效果越差(图 2(a)，分发完成之后总的副本污染率仅为 1007/8007=12.6%)；相反，较低的警觉度会造成大量的额外流量(图 2(b)，总的副本污染率高达 6545/13545=48.3%)并使分发完成的时间延长(图 2(b)，完成时间加倍)，甚至使分发过程失败(图 2(c))。



(a)totally 8007 Downloads, 1007 bad

(b)totally 13545 Downloads, 6545 bad

(c)totally 5487 Downloads, 5363 bad

图 2 警觉度对文件污染效果的影响

取 nAwareMax=nUpMax/2 这个警觉度参数(图 2(b))来分析本文提出的“提醒-删除”机制的污染防治效果。如果采用了这一机制，则：每个下载者记录自己的副本来源节点，若发现版本是坏的，则在删除副本、插入下载队列之前，还向其来源节点发出一个警告信息；被警告的节点会提前(在不足A_p轮的时候)发现自己的版本错误，从而停止上传，截断来自自己的这个污染源。实验结果如图 3 所示。

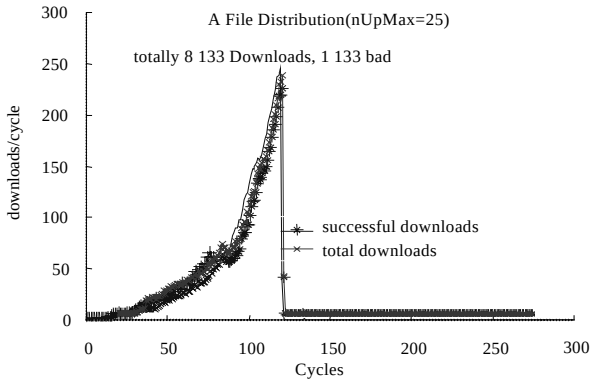


图 3　版本污染与“提醒-删除”对策

对比图 3 与图 2(b)的两组曲线可以看出，“提醒-删除”机制可以有效地降低文件污染的成功率，总的副本污染率由 48.3%降至 13.9%，耗时减半，几乎接近于图 2(a)中 nAwareMax=0 的理想情况(事实上，nAwareMax=0 时，版本污染已经退化为副本污染)。

4　结语

本文分析了 P2P 网络中文件污染的 3 种主要方式(版本污染、副本污染和索引污染)，并提出了相应的对策。针对其中使用最多和危害性最大的版本污染，提出了一种新的“提醒-删除”对策。借助仿真实验揭示了 P2P 文件分发与文件污染的规律，表明节点“慷慨度”的高低对文件分发效率所起到的决定性的作用，以及节点“警觉度”对文件污染(特指版本污染)效果的影响。

实验结果表明，“提醒-删除”机制可以有效地提高节点的整体警觉度，增加污染的难度和降低污染的成功率，从而在一定程度上缓解 P2P 网络中严峻的文件污染问题。

参考文献

1 Christin N, Weigend A S, Chuang J. Content Availability, Pollution and Poisoning in Peer to Peer File Sharing Networks[C]//Proc. of ACM E-Commerce Conference. 2005-06: 68-77.

2 Walsh K, Sirer E G. Fighting Peer to Peer SPAM and Decoys with Object Reputation[C]//Proc. of ACM SIGCOMM’05 Workshops. 2005-08: 138-143.

3 Lee U, Choi M, Cho J, et al. Understanding Pollution Dynamics in P2P File Sharing[C]//Proc. of IPTPS’06. 2006-02.

4 Liang J, Naoumov N, Ross K W. The Index Poisoning Attack in P2P File-sharing Systems[C]//Proc. of INFOCOM’06. 2006-04.

5 Liang J, Kumar R, Xi Y, et al. Pollution in P2P File Sharing Systems[C]//Proc. of INFOCOM’05. 2005-03: 1174-1185.

(上接第 21 页)

3　安全分析

3.1　系统公/私密钥的安全性

本协议采用无需可信秘密分发者的可验证秘密(k,n)分享方案产生系统公/私密钥，系统私人密钥不会在协议中显式出现，必须有门限个管理者才能恢复密钥和签发证书，攻击者必须攻击 k 个管理者才能获取系统密钥。为防止攻击者对系统进行长期攻击，系统的密钥可定期更新，以达到更高的安全性。

3.2　消息重发攻击

完全分布式系统中通信双方每次会话都可能被窃听并被记录，窃听者会在以后的会话中重放这些协议[5]，为了保持协议消息的新鲜性，通信双方在数据中增加时间戳，这样在通信中可通过检查时间戳发现重放攻击。

此外，应答消息中的$P_j'(TS)$和$Cert_j$中的B_j'可防止其他peer使用$Cert_j$来伪装peer j提供假的信任值。

3.3　消息伪造攻击

若THA peer提供虚假信任值，则可通过$Cert_j$定位恶意THA peer；非THA peer无法产生合法的应答消息，因为只有THA peer才知道 SP_i；交互双方可通过查询对方信任值得到对方的公开密钥，因此在交换评估证据时无法伪造密钥对$\langle P,B\rangle$；只有THA peer知道SP_i，因此评估报告在传输中不可能被篡改。

3.4　匿名性

系统中保持匿名性可以保护各 peer 免遭恶意 peer 的攻击。在应答消息中包含的是 THA 证书而不是 ID，既可证明自己的合法身份又保持了匿名性；在评估报告消息中包含交互证据而不出现用户 ID，恶意用户无法知道是谁提交了“差评”报告。

4　结论

信任关系管理是信任模型的重要内容，在笔者提交的管理协议中，利用基于门限机制的认证技术，为 P2P 分布式的网络提供证书服务，系统中无需可信中心节点，提高了系统的可靠性、安全性；管理协议提供匿名的信任查询服务，经分析，协议可提供安全、可靠和可追究责任的信任管理服务。

参考文献

1 Aberer K, Despotovic Z. Managing Trust in a Peer-to-Peer Information System[C]//Proceedings of the 10th International Conference on Information and Knowledge Management, New York, USA. 2001: 310-317.

2 Singh A, Liu L. TrustMe: Anonymous Management of Trust Relationships in Decentralized P2P Systems[C]//Proceedings of IEEE International Conference on P2P Computing, Sweden. 2003: 142-149.

3 Shamir A. How to Share a Secret[J]. Communications of the ACM, 1979, 22(11): 612-613.

4 Pedersen T P. A Threshold Cryptosystem Without a Trusted Party[C]// Proc. of Eurocrypt’91. [S.l.]: Springer-Verlag, 1991: 522-526.

5 Aura T. Strategies Against Replay Attacks[C]//Proc. of Computer Security Foundations Workshop. 1997: 59-68.