韓國컴퓨터情報學會 論文誌
第16卷 第3號, 2011. 3.

2011-16-3-3-2

DHT 기반의 P2P 네트워크에서 사용자 행동양식 및 파일 오염에 관한 측정 연구

신 규 용*, 유 진 철*, 이 종 덕*

A Measurement Study of User Behavior and File Pollution in DHT-based P2P Networks

Kyuyong Shin*, Jincheol Yoo*, Jongdeog Lee*

요 약

부패한 파일을 공유하거나, 인덱스 정보에 잘못된 인덱스 레코드를 삽입하는 등의 파일 오염문제는 대다수의 파일 공유 P2P 시스템들의 실질적인 문제가 되어 왔다. 이러한 파일 오염은 사용자들로 하여금 다운받아도 전혀 득이 되지 않는 오염된 파일들을 다운받도록 하거나, 존재하지 않는 파일들에 대한 비생산적인 다운로드 시도를 유도한다. 파일 오염은 네트워크 자원을 낭비할 뿐만 아니라, 사용자들의 활발한 참여를 제한하기 때문에 적절하게 대처하지 못한다면 향후 파일 공유 P2P 시스템 (혹은 비슷한 분산 환경 정보 공유 어플리케이션)의 성공을 기약하기 힘들다. 따라서 효과적인 오염방지 메커니즘의 개발이 시급하다. 본 논문은 대표적인 DHT (distributed hash table) 기반 P2P 시스템인 Kad 네트워크에서 사용자 행동양식 및 파일 오염에 대한 측정 연구를 통해 향후 효과적인 파일 오염 방지 메커니즘을 개발하고자 하는 연구자들에게 실질적으로 활용 가능한 정보를 제공한다.

▸ Keyword : 피투피, 파일 공유, 분산시스템, 파일 오염, 오염 방지

Abstract

File pollution (i.e., sharing of corrupted files, or contaminating index information with bogus index records) is a de facto problem in many file sharing Peer-to-Peer (P2P) systems in use today. Since pollution squanders network resources and frustrates users with unprofitable downloads (due to corrupted files) and unproductive download trials (due to bogus index records), the viability of P2P systems (and similar distributed information-sharing applications) is questionable unless properly addressed. Thus, developing effective anti-pollution mechanisms is an immediate problem in this literature. This paper provides useful information and deep insight with future researchers who want to design an effective anti-pollution mechanism throughout an extensive measurement study of user behavior and file pollution in a representative DHT-based P2P system, the Kad network.

▸ Keyword : Peer-to-Peer, File Sharing, Distributed System, File Pollution, Anti-pollution

• 제1저자, 교신저자 : 신규용
• 투고일 : 2010. 10. 25, 심사일 : 2010. 12. 01, 게재확정일 : 2010. 12. 06.
* 육군사관학교 전자정보학과 (Dept. of Electrical Engineering and Information Science, Korea Military Academy)
※ 본 논문은 육군사관학교 화랑대연구소 2011년도 연구 활동비를 지원받아 연구되었음

Ⅰ. 서 론

최근 Peer-to-Peer (P2P) 기반의 파일 공유 시스템이 기존의 클라이언트-서버 구조를 대신하는 새로운 대안으로 각광받고 있다. 하지만 대부분의 파일 공유 P2P 시스템들은 초기 시스템 설계 단계에 파일 오염이라는 문제에 대한 고려가 없었기 때문에 의도적인 파일 오염 공격에 취약하다[1]. 그 결과 이미 많은 파일 공유 P2P 시스템에서 파일 오염이 심각한 수준에 이르렀고, 그로 인해 네트워크의 자원 및 사용자들의 시간이 헛되이 낭비되고 있다[1,2,3]. 따라서 효과적인 파일 오염 방지기법의 개발은 향후 파일 공유 P2P 시스템의 성공에 필수적인 전제조건이라 하겠다.

효과적인 파일 오염 방지기법을 개발하기 위해서는 먼저 파일 공유 P2P 시스템 사용자들의 행동양식 및 파일 오염 공격에 대한 실태분석이 선행되어야 한다. 이를 위해 본 논문은 현존하는 DHT 기반 파일 공유 P2P 시스템들 중에서 순간 사용자가 백만이 넘을 정도로 그 규모가 가장 큰 Kad 네트워크를 대상으로 측정연구를 수행한다. 데이터 수집을 위해서 가장 대표적인 Kad 네트워크 접속 클라이언트인 eMule (0.49a MorphXT version 11.0[4])을 수정해 측정노드 (crawler)를 개발하였다. 개발된 측정노드는 Kad 네트워크에 참여해 파일 및 출판자에 대한 출판 메시지를 수신할 수 있도록 설정되었다. 측정노드를 통해 수집된 출판 메시지들에 대한 분석을 토대로 본 논문은 Kad 네트워크 사용자들의 파일 공유 패턴, 공유되는 파일들의 버전 인기도, 사용자들의 사설 IP 주소 사용 비율, IP 주소 프리픽스 (prefix) 범위별 동일 버전 다운로더들의 수 등을 분석하였다. 나아가 측정노드에 메시지 내용 검증 절차를 추가해 현재 Kad 네트워크 내에서의 파일 오염 정도를 인덱스 오염 중심으로 확인하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 DHT 기반 파일 공유 P2P 시스템에 대한 이해를 돕기 위해 이 분야에서 사용되는 기본용어들을 소개하고, DHT를 이용한 파일의 출판 및 검색 과정을 설명한다. Ⅲ장에서는 대표적인 파일 오염 공격방법들에 대해 소개하고, Ⅳ장에서는 측정 환경 및 방법에 대해 자세히 기술한다. Ⅴ장에서는 수집된 결과를 분석하고, Ⅵ장에서는 본 논문과 연관된 관련 연구들을 소개한다. 마지막으로, Ⅶ장에서는 결론 및 향후 연구방향에 대해 서술한다.

Ⅱ. 배경지식

본장에서는 먼저 DHT 기반 파일 공유 P2P 시스템에 대한 이해를 돕기 위해 이 분야에서 사용되는 기본적인 용어들을 소개하고, DHT를 이용한 파일의 출판 및 검색 과정에 대해 Kad 네트워크를 기준으로 자세히 설명한다.

1. 파일 공유 시스템에서 사용되는 기본용어

파일 공유 시스템에서 공유되는 영화나 음악 파일과 같은 특정 파일을 타이틀 (title)이라 부르고, 하나의 타이틀은 생성 방법에 따라 여러 가지 버전 (version)이 있을 수 있다. 만일 하나의 버전을 여러 사용자들이 다운로드해 재 공유한다면 시스템 내에 여러 개의 복사본 (copies)이 존재할 수 있다. 파일 오염과 관련해서 실제로는 존재하지 않는 파일이거나, 오염된 (혹은 질이 낮은) 내용물을 담고 있는 미끼 (decoy)가 존재해 사용자들을 현혹시킨다. 각각의 파일들은 파일 이름, 크기, 형태 및 포맷, 기타 정보가 기록된 메타데이터 (meta-data)를 포함하고 있다[2,5]. 파일의 메타데이터 (주로 파일 이름)로부터 추출된 하나의 토큰 (token)은 키워드 (keyword)라 불리며, Kad 네트워크에서 사용되는 키워드는 주로 3개 이상의 영문자로 구성된다[5]. 따라서 파일 이름이 여러 단어로 구성되어 있는 파일의 경우 다수의 키워드를 포함할 수 있다. 사용자 (user)는 P2P 클라이언트 프로그램을 운영하는 사람이고, 피어 (peer) 혹은 노드 (node)는 클라이언트 프로그램 자체를 지칭한다. P2P 시스템 내에서 파일을 공유하는 사람을 콘텐츠 소유자 (content owner) 혹은 출판자 (publisher)라 부르고, 파일을 다운로드 하는 사람은 다운로더 (downloader)라 부른다.

2. DHT를 통한 출판 및 검색 과정

Kad 네트워크는 eMule[6] 또는 aMule[7] 등 다양한 형태의 파일 공유 어플리케이션으로 구현되어 사용되어지고 있는 대표적인 DHT 기반 파일 공유 P2P 시스템이다. Kad 네트워크는 가장 잘 알려진 DHT 중의 하나인 Kademlia[8]에 기반하고 있으며, 전 세계적으로 매 순간 동시 사용자가 백만 명이 넘는다[9]. Kad 네트워크에서 DHT에 저장되는 인덱스 정보는 키 (key)로 구분되며, 각각의 키는 사용자 아이디 (ID)와 동일한 형태 및 비트 수를 갖는다. 출판과 검색을 위해서 콘텐츠 키 (content key)와 키워드 키 (keyword key) 등 두 가지 형태의 키가 이용된다. 콘텐츠 키는 파일 내용 전체에 대한 해시 (hash) 값으로, 동일한 버전의 파일들은 동일한 콘텐츠 키 값

(a) 출판 (b) 검색

그림 1. DHT 기반 P2P 시스템에서 출판과 검색의 개요
Fig. 1. A high level overview of the publish and retrieve mechanism in a DHT-based P2P system

을 갖는다. 키워드 키는 해당 파일의 메타데이터 (주로 파일 이름)의 각 키워드에 대한 해시 값이다.

출판은 특정 파일 및 출판자에 대한 정보를 DHT에 저장하는 과정이다. Kad 네트워크에서 각각의 인덱스 노드들은 자신의 노드 아이디 주변의 특정한 부분을 담당하여 파일 정보를 저장한다. 콘텐츠 소유자, 즉 파일을 출판하고자 하는 출판자는 먼저 출판하고자 하는 파일 내용 전체를 해시하여 콘텐츠 키를 얻은 후, 얻어진 콘텐츠 키와 가장 가까운 Kad 아이디를 갖는 최기 인덱스 노드를 찾는다. 현재 eMule은 하나의 키와 관련된 정보를 저장하기 위해 최기 인덱스 노드뿐만 아니라 최기 인덱스 노드와 인접한 다수의 인덱스 노드들 (tolerance zone)을 찾아 정보를 저장하는, 다소 모호하고 복잡한 알고리즘을 사용한다. 그러나 이는 출판과 검색 과정의 이해에 영향을 미치지 않으므로 본 논문에서는 출판자가 최기 인덱스 노드만을 찾는다고 가정한다. 최기 인덱스 노드를 찾기 위해서 출판자는 반복적 라우팅 (iterative routing) 알고리즘[5,10]을 활용한다. 출판자가 최기 인덱스 노드를 찾고 나면, 그 인덱스 노드는 출판되는 파일의 콘텐츠 키 소유자가 된다. 콘텐츠 키 소유자는 자신의 콘텐츠 인덱스 테이블에 출판되는 파일에 대한 <콘텐츠 키, 출판자 정보>를 저장한다. 이때 출판자 정보는 출판자의 Kad 아이디, IP 주소, 포트 번호 등을 포함한다. 이와 같이 출판자가 콘텐츠 키 소유자를 찾아 출판자 자신의 정보를 저장하는 과정을 콘텐츠 출판이라 부른다. 다음으로 출판자는 출판하는 파일의 메타데이터에서 추출된 각각의 키워드를 해시하여 키워드 키를 얻은 다음, 앞서 설명한 것과 동일한 방법으로 각 키워드 키에 대한 최기 인덱스 노드인 키워드 키 소유자를 찾는다. 각각의 키워드 키 소유자는 자신의 키워드 인덱스 테이블에 출판되는 파일에 대한 <키워드 키, 콘텐츠 키, IP 주소 목록, 가용한 파일이름, 메타데이터>를 저장한다. 이와 같이 출판자가 출판하는 파일의 각 키워드 키 소유자를 찾아 파일에 관한 정보를 저장하는 과정을 키워드 출판이라 부른다. 콘텐츠 출판에 이은 키워드 출판 형태를 2단계 출판 스킴이라 부른다[5]. Kad 네트워크에서는 사용자들이 수시로 바뀌는 것을 수용하기 위해 콘텐츠 출판은 매 5시간마다, 그리고 키워드 출판은 매 24시간마다 반복된다.

DHT에서 원하는 파일에 대한 검색은 출판의 역순이다. 파일 다운로드를 원하는 사용자 (다운로더)는 유추되는 파일의 이름에서 가장 적합한 키워드를 선택한 후 해시하여 키워드 키를 계산한다. 계산된 키워드 키를 바탕으로 반복적 라우팅을 통해 최기 인덱스 노드인 키워드 키 소유자를 찾아 콘텐츠 키 및 메타데이터 목록을 얻은 후, 메타데이터 정보를 바탕으로 자신이 원하는 파일과 가장 부합되는 콘텐츠 키를 선택한다. 이와 같이 키워드를 통해 원하는 콘텐츠 키를 찾는 과정을 키워드 검색이라 부른다. 선택된 콘텐츠 키는 다시 출판자 정보를 저장하고 있는 최기 인덱스 노드인 콘텐츠 키 소유자를 찾는데 이용되며, 찾아진 콘텐츠 키 소유자로부터 출판자에 대한 정보 (IP 주소 및 포트 번호 등)를 얻는다. 이렇게 콘텐츠 키를 바탕으로 출판자에 관한 정보를 찾아가는 과정을 콘텐츠 검색이라 부른다. 마지막으로 다운로더는 얻어진 출판자 정보를 바탕으로 출판자와 직접 연결해 원하는 파일을 다운로드한다. 다운로드 된 파일은 즉시 자동으로 재 출판되며, 다운로더가 인위적으로 해당 파일을 지우거나 재 공유되지 않도록 설정하지 않는 한 계속 공유된다.

그림 1은 DHT 기반 P2P 시스템에서 출판 및 검색 과정을 예시하고 있다. 이 예에서 출판자는 "Dragon War.mpg"라는 영화파일을 출판하고, 다운로더는 동일한 파일을 다운

로드한다고 가정한다. 그림 1(a)에서 보듯이, 파일을 출판하기 위해 출판자는 파일 내용 전체를 해시하여 콘텐츠 키 (ck)를 얻고, 콘텐츠 출판 메시지를 콘텐츠 키 소유자에게 보내 출판자 정보를 저장하도록 한다. 다음으로 각 키워드 ("dragon"과 "war")를 해시하여 키워드 (kk1, kk2) 키를 얻고, 키워드 출판 메시지를 각 키워드 키 소유자에게 보내 출판되는 파일에 대한 정보 (콘텐츠 키 및 메타데이터)를 저장하도록 한다. 해당 파일을 다운로드하기 위해 다운로더는 그림 1(b)에서 보듯이 메타데이터에 포함되어 있을 법한 키워드 (이 예제에서는 "dragon")를 선택하여 해시한다. 이렇게 얻어진 키워드 키를 기반으로 키워드 탐색 메시지를 키워드 키 소유자에게 보내 원하는 파일에 대한 콘텐츠 키 및 메타데이터 목록을 얻는다. 다음으로 키워드 키 소유자로부터 얻어진 콘텐츠 키 목록에서 (메타데이터를 바탕으로) 하나의 콘텐츠 키를 선택한다. 다운로더는 선택된 콘텐츠 키를 이용해 콘텐츠 탐색 메시지를 콘텐츠 키 소유자에게 보내 출판자에 대한 정보를 얻는다. 마지막으로 출판자와 직접 연결한 뒤 해당 파일을 다운로드 받는다.

III. 현존하는 파일 오염 공격들

일반적으로 파일 오염 공격은 공격자 (polluter)들이 채택하는 방법들에 따라 크게 콘텐츠 오염, 메타데이터 오염, 그리고 인덱스 오염 등으로 구분된다.

1. 콘텐츠 오염 (content pollution)

콘텐츠 오염[2,11,12]은 목적파일의 내용 일부 혹은 전체를 바꾸어 버림으로써 파일의 질을 떨어뜨리는 공격 방법이다. 이러한 공격은 파일 내용에 노이즈 또는 쓰레기를 삽입, 내용을 일부를 생략하거나 내용의 순서를 뒤섞어 놓기, 혹은 내용 전체를 전혀 새로운 것으로 바꾸기 등의 방식으로 이루어진다. 공격자들은 오염되지 않은 목적파일과 동일한 콘텐츠 키 값을 갖는 오염된 파일들을 쉽게 만들 수 있는데, 이는 Kad 네트워크에서 사용되는 해시 함수인 MD4[13]의 취약성을 이용한 것이다. 콘텐츠 오염 방법은 크래커 (cracker)들이 파일 공유 P2P 시스템을 통해 바이러스를 유포할 때 사용하는 대표적인 방법이다.

2. 메타데이터 오염 (meta-data pollution)

메타데이터 오염[1,2]은 파일의 내용 자체에 대한 공격이라기보다는 파일에 대한 기본 정보를 담고 있는 메타데이터에 대한 공격이다. 앞서 II-2장에서 보듯이, 다운로더들은 키워드 탐색을 통해 얻어진 콘텐츠 키 목록에서 하나의 콘텐츠 키를 선택할 때 메타데이터의 내용을 참조한다. 따라서 메타데이터의 내용이 파일의 내용을 제대로 묘사하지 못하고 있다면 다운로더가 자신이 원하지 않는, 전혀 다른 파일을 다운받을 수 있다. 특정 파일의 이름을 다른 이름으로 바꾸는 것은 메타데이터 오염의 대표적인 예라 할 수 있다. 예를 들어 그림 1(a)에서 공격자가 "Haeundae.mpg"라는 영화파일의 제목을 "Dragon War.mpg"라고 바꾸어 출판한다면, "Dragon"이라는 키워드를 사용해 "Dragon War.mpg"를 탐색하는 다운로더들은 이름은 비록 "Dragon War.mpg"이지만 실제 내용은 "Haeundae.mpg"인 영화파일을 다운받을 수도 있다.

3. 인덱스 오염 (index pollution or poisoning)

인덱스 오염[3]은 파일의 내용이나 메타데이터를 공격하기보다는 DHT의 인덱스 구조를 직접 공략한다. 이러한 인덱스 오염은 다량의 가짜 키워드 혹은 콘텐츠 출판 메시지들을 콘텐츠 키 소유자 혹은 키워드 키 소유자에게 보냄으로써 인덱스 노드들이 잘못된 파일 혹은 출판자 정보를 저장하도록 유도한다. 앞서 그림 1(b)에서 보듯이, 다운로더들이 원하는 파일을 찾기 위해서는 인덱스 노드 (콘텐츠 키 소유자와 키워드 키 소유자)들이 제공하는 파일 및 출판자 정보에 의존할 수밖에 없다. 따라서 대다수의 인덱스 노드들이 잘못된 인덱스 정보를 저장하고 있다면 다운로더들은 인덱스 노드들이 제공하는 정보로부터 자신들이 원하는 파일을 전혀 찾지 못하거나 찾게 되더라도 많은 시간을 허비하게 된다. 대부분의 파일 공유 P2P 시스템들에서 인덱스 노드들은 출판 메시지의 내용이나 출판자에 대한 검증을 하지 않기 때문에, 오염 공격자들은 인덱스 노드들이 가지고 있는 파일 및 출판자 정보를 쉽게 오염시킬 수 있다. 인덱스 오염은 현재 가장 보편적인 공격방법으로 알려져 있다[3].

IV. 측정 환경 및 방법

실험에 사용된 측정노드는 북미 노스캐롤라이나 주립대학교 (North Carolina State University)에 위치하며, Kad 네트워크에 직접 접속해 측정을 실시한다. 실험을 위해 Top 10 Songs[14] 사이트에서 가장 인기도가 높은, 그래서 파일 오염 공격의 대상이 되기 쉬운 4개의 mp3 음악파일들[2]을 선택하였

1) 본 논문에서는 저작권 문제로 실험에 사용된 음악파일들에 대한 정확한 이름을 사용하는 대신 T_1 ~ T_4로 단순화하여 명시하였다.

다. 또한 비교목적으로 1970년대 빌보드차트에서 가장 인기가 있었던 곡들 중 한 곡에 대한 mp3 음악파일을 추가로 선택하였는데, 선택된 곡 (T_5)은 현재는 별로 유명하지도 않고, 저작권 문제도 없기 때문에 파일 오염공격에 대한 위협이 상대적으로 적을 것으로 판단되었다.

1. 사용자 행동양식에 대한 데이터 수집방법

사용자 행동양식 연구에 필요한 데이터 수집을 위해, 현존하는 eMule 클라이언트 (0.49a MorphXT version 11.0[4])를 수정하여 Kad 네트워크 용 측정노드 (crawler)를 제작하였다. 측정노드가 각각의 음악파일에 대한 키워드 출판 메시지를 수집할 수 있도록 하기 위해, 즉 각 음악파일에 대한 키워드 키 소유자가 될 수 있도록, 각 곡의 이름에서 대표적인 키워드 (K_{T_i}) 하나를 선택한 후 128 비트 키워드 키로 해시하였다. 이후 측정노드의 Kad 아이디를 얻어진 각각의 키워드 키와 동일하게 설정함으로써 측정노드로 하여금 같은 키워드를 사용하는 파일들에 대한 키워드 출판 메시지를 수집할 수 있도록 하였다. 다음으로, 콘텐츠 출판 메시지에 대한 정보를 수집하기 위해 수집된 키워드 출판 메시지에 대한 분석을 통해 얻어진 각 음악파일의 콘텐츠 키들 중에서 가장 빈도수가 높은 콘텐츠 키, 즉 가장 유명한 버전을 선택하여 해시한 후 측정노드의 Kad 아이디로 사용하였다. 이렇게 함으로써 측정노드가 해당 콘텐츠 키에 대한 콘텐츠 키 소유자가 되어 각 음악파일의 최고 인기버전에 대한 콘텐츠 출판 메시지를 수집할 수 있게 된다. 이렇게 수집된 키워드 및 콘텐츠 출판 메시지들에 대한 분석을 통해 Kad 사용자들의 행동양식을 분석하였고, 그 결과는 Ⅴ-1장에서 확인할 수 있다.

2. 파일 오염에 대한 데이터 수집방법

Kad 네트워크에서 파일 오염에 대한 데이터 수집을 위해 앞서 사용자 행동양식에 대한 데이터 수집에 사용되었던 측정노드의 기능이 일부 보완되었다. 즉, 앞서 사용된 측정노드는 다른 보통의 eMule 클라이언트와 마찬가지로 (키워드 및 콘텐츠) 출판 메시지에 대한 아무런 검증 없이 바로 인덱스 테이블에 해당 정보를 저장한다. 그러나 보완된 측정노드는 파일 오염에 대한 데이터를 수집하기 위해 출판 메시지를 받을 때마다 (1) 정당한 사용자로부터 보내진 출판 메시지인지, (2) 출판된 메시지의 내용이 올바른 지를 확인하는 기능이 추가되었다. 정당한 사용자로부터의 출판 메시지인지를 점검하기 위하여 보완된 측정노드는 출판 메시지를 받을 때마다 어플리케이션 레벨의 ping 메시지를 보낸다[2]. 이때 출판노드로부터 어플리케이션 레벨 pong 메시지가 도착한다면 정당한 사용자로부터의 출판이라 판단한다. 이러한 확인을 통해 하나의 파일 오염 공격자가 주소 도용 (IP spoofing)을 통해 다수의 거짓 출판 메시지를 보내는 것을 방지할 수 있다. 출판된 메시지의 내용이 올바른 지 확인하기 위해 보완된 측정노드는 첫 번째 단계를 통해 정당한 사용자로부터의 출판 메시지임이 확인된 메시지에 대해서만 그 내용을 확인하는 절차를 거친다. 먼저 보완된 측정노드는 키워드 출판 메시지를 받을 때마다 메시지에 포함된 콘텐츠 키를 바탕으로 콘텐츠 탐색을 실시하여 45초[5] 이내에 출판자 정보를 얻을 수 있는 지를 확인한다. 만일 시간 안에 출판자에 대한 정보를 얻을 수 있으면 그 출판 메시지는 진짜 (genuine)로 분류하고, 동일한 콘텐츠 키를 포함하는 키워드 출판 메시지에 대해서는 추가적인 검증을 실시하지 않는다. 하지만 주어진 시간 내에 출판자 정보를 얻지 못하면 그 키워드 출판 메시지는 가짜 (bogus)로 분류되어 버려진다. 다음으로 콘텐츠 출판 메시지에 대해서는 콘텐츠 출판 메시지에 포함된 출판자 정보 (IP 주소와 포트번호)를 통해 출판자와의 TCP 연결이 가능한 지 확인한다. 만일 TCP 연결이 불가능하다면 해당 콘텐츠 출판 메시지는 가짜로 분류하고 버린다. 이와 같은 단계를 통해 보완된 측정노드는 가짜 출판 메시지로부터 얻어진 파일 및 출판자 정보가 자신의 인덱스 테이블에 저장되는 것을 방지할 수 있다. 실험간 키워드 출판 메시지에 대해서는 48시간 동안, 콘텐츠 출판 메시지에 대해서는 10시간 동안 데이터를 수집했는데 이는 일반적인 재출판 주기의 2배에 해당한다. 이렇게 수집된 데이터를 바탕으로 현재 Kad 네트워크 안에서 인덱스 오염 정도를 확인할 수 있었으며 그 결과는 Ⅴ-2장에서 확인할 수 있다. 측정과 관련된 보다 자세한 내용은 본 논문의 확장 버전인 NCSU 연구보고서[15]를 참조할 수 있다.

Ⅴ. 측정결과 분석

본 장에서는 먼저 사용자 행동양식에 대한 데이터 수집방법 (Ⅳ-1장)을 통해 수집된 출판 메시지들을 분석하여 DHT 기반의 파일 공유 P2P 시스템 사용자들의 행동양식을 분석한다. 다음으로 파일 오염에 대한 데이터 수집방법 (Ⅳ-2장)을 통해 수집된 데이터를 바탕으로 Kad 네트워크에서 파일 오염

2) eMule의 경우 클라이언트간의 연결을 보장하기 위해 HELLO 메시지를 구현하고 있으므로 이 메시지를 이용해 간단하게 어플리케이션 레벨의 ping/pong을 구현할 수 있었다. 물론 ICMP 메시지를 이용한 ping을 통해 동일한 기능을 구현할 수 있지만 실험 결과 많은 네트워크에서 해킹 방지를 위해 ICMP 메시지를 차단하고 있어 정확한 통계를 얻기에는 부적합하였다.

도를 인덱스 오염을 중심으로 측정한다. 이렇게 얻어진 일련의 측정 및 분석결과는 앞으로 효과적인 파일 오염방지 메커니즘을 개발하고자 하는 연구자들에게 유용하게 활용될 수 있을 것이다.

1. 사용자 행동양식에 대한 측정결과 분석

먼저 본 논문은 Kad 네트워크에서 출판되는 키워드 및 콘텐츠 출판 메시지들에 대한 분석을 통해 출판 메시지들의 분포, 공유되는 파일들의 버전 인기도, 사용자들의 사설 IP 주소 사용 비율 및 동일 버전 다운로더들의 IP 주소 분포 등을 살펴본다.

그림 2는 각 키워드 키 소유자에 의해서 수신된 5개의 mp3 음악파일들에 대한 시간당 전체 키워드 출판 메시지 개수를 보여준다. 결과에서 보듯이 키워드 출판 메시지 개수는 시간과 매우 밀접한 관련이 있음을 보여준다. 또한 각각의 파일에 사용되는 키워드의 인기도에 따라 각 인덱스 노드 (키워드 키 소유자)가 처리해야 하는 부담이 수십 배 이상 차이나고 있음을 볼 수 있다. 측정노드가 북미지역인 노스캐롤라이나 주립대학에 위치하고 있다는 점과 측정에 사용된 음악파일들이 미국에서 유행하고 있는 팝송들임을 감안할 때 Kad 사용자들의 경우 주로 주간에 활동이 왕성함을 유추할 수 있다.

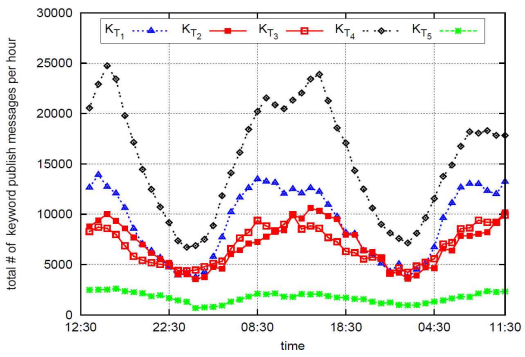

그림 2. 시간당 수신된 전체 키워드 출판 메시지 개수
Fig. 2. Total # of keyword publish messages per hour

그림 2의 결과에서 보듯이, 하나의 키워드 키 소유자가 처리해야 하는 키워드 출판 메시지의 개수는 상당히 많다. 하지만 사용자들의 다운로드에 의해 한 버전에 대한 복사본 (copy) 수가 증가할수록 동일한 콘텐츠 키 (버전)를 포함하는 키워드 출판메시지가 계속 출판된다는 점과, 공유되는 모든 파일이 주기적으로 재출판 된다는 점을 감안할 때 이전에 출판되지 않았던 새로운 버전에 대한 키워드 출판 메시지 개수는 상대적으로 적을 것으로 예상되었다. 이를 확인하기 위해, 서로 다른 콘텐츠 키 (즉 버전)를 포함하고 있는 키워드 출판 메시지만의 개수가 조사되었고, 그림 3은 그 결과를 보여준다. 이 결과에서 보듯이 대부분의 경우 이전에 출판된 적이 없는 새로운 버전에 대한 키워드 출판 메시지의 개수는 전체 키워드 출판 메시지의 1/5 이하로 적고, 시간이 지남에 따라 그 비율도 줄어들고 있음을 보여준다. 결론적으로 현존하는 파일 공유 P2P 시스템 내에서는 새로운 파일의 출판보다는 이미 존재하는 파일들의 확산이 지배적임을 알 수 있다.

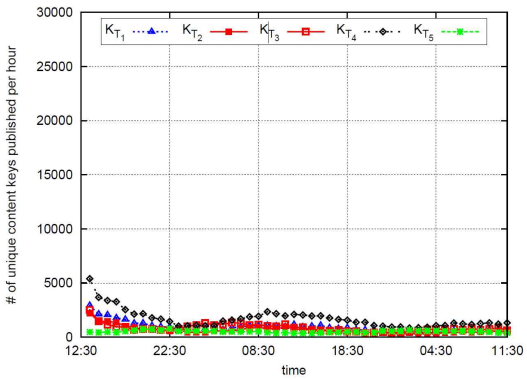

그림 3. 새로운 버전에 대한 출판 메시지 개수
Fig. 3. Total # of unique content keys

다음으로 각 음악파일에 대한 버전 인기도를 조사하였다. Kad 네트워크에서는 동일한 콘텐츠 키를 포함하는 키워드 출판 메시지 개수가 곧 그 버전에 대한 인기도를 의미한다. 왜냐하면 Kad 네트워크에서는 하나의 파일이 다운로드 될 때마다 그 파일의 콘텐츠 키를 포함하는 키워드 출판 메시지가 자동으로 발행되며, 동일한 버전의 파일은 동일한 콘텐츠 키값을 갖기 때문이다. 인기도를 조사하기 위해 측정에 사용된 음악파일들의 각 버전을 측정 기간 동안 수신된 키워드 출판 메시지의 개수 순으로 정렬하였다. 그림 4는 각 음악파일의 가장 유명한 100개의 버전에 대한 로그-로그 스케일의 확률밀도함수 (Probability Density Function, PDF)를 보여준다. 직선에 가까운 선들은 측정에 사용된 음악파일들의 버전 인기도가 지프 분포 (Zipf distribution)를 따름을 알 수 있다. 이는 Kad 사용자들이 다른 선택의 기준이 명시되지 않는 경우 각 버전의 인기도를 기반으로 다운로드할 버전을 선택함을 의미한다. 이 결과는 다른 P2P 기반의 파일 공유 시스템들에 대한 선행연구들[2,16]의 결과와 일치한다.

파일 공유 P2P 시스템에서 NAT (Network Address

Translation) 서버 혹은 방화벽 (fire-wall) 뒤에 위치하는 사설 IP 주소 사용자[2]의 분포는 효과적인 파일 오염방지 시스템 설계에 있어서 중요한 고려요소 중의 하나이다. 따라서 이번 실험에서는 Kad 네트워크에서 사설 IP 주소를 사용하는 사용자의 분포를 조사한다. 사설 IP 주소의 분포는 파일을 다운로드하는 사용자들의 지역적 분포를 이해하여 효과적인 평판 시스템 (reputation system)을 설계하는데 도움을 준다. Kad 네트워크에서 사설 IP 주소를 사용하는 출판자의 경우 콘텐츠 출판 메시지에 공인 IP 주소를 가지는 동료 (buddy) 정보를 포함하게 되는데, 그 이유는 다운로더들이 사설 IP 주소를 사용하는 출판자에게 바로 TCP 연결을 할 수 없기 때문이다. 이런 경우 출판자는 콘텐츠 출판 메시지에 동료 정보를 제공하여 다운로더들이 그 동료를 통해 우회적으로 연결을 할 수 있도록 유도한다. 따라서 콘텐츠 키 소유자들은 콘텐츠 출판 메시지에 포함된 동료 정보의 유무를 통해 출판자가 사설 IP 주소 사용자인지를 쉽게 판단할 수 있다.

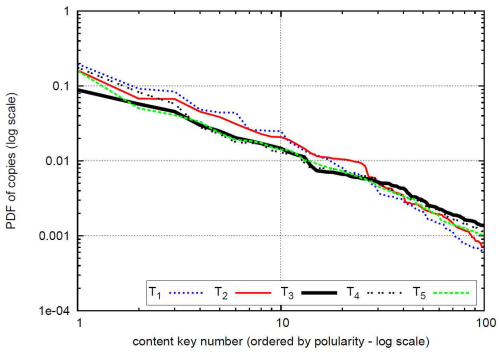

그림 4. 음악파일 별 상위 100개 버전에 대한 인기도
Fig. 4. Version popularity of top 100 content keys per title

그림 5는 측정에 사용된 각 음악파일들의 최고 인기버전을 다운로드 하는 사용자들에 대한 사설 IP 주소 분포를 보여준다. 결과에서 보듯이 사설 IP 주소를 사용하는 다운로더들의 비율은 전체 사용자의 38.55%에서 46.06%에 이른다. 본 실험에서 사용된 음악파일들이 공인 IP 주소가 가장 많이 할당된 북미지역 (ARIN)에서 인기 있는 곡들이라는 점을 감안한다면 실제 사설 IP 주소 사용자의 비율은 이보다 훨씬 높을 것으로 판단된다.

3) 엄밀히 말해 NAT 서버나 방화벽 뒤에 위치하는 사용자가 모두 사설 IP 주소를 사용한다고 볼 수는 없지만, 본 논문에서는 자신의 IP 주소가 아닌 다른 공인 IP 주소를 사용하는 사용자를 모두 사설 IP 주소 사용자로 분류한다.

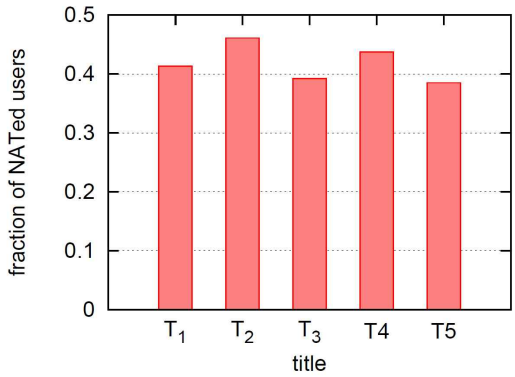

그림 5. 사설 IP 주소 분포
Fig. 5. Distribution of private IP addresses (Kad)

마지막으로, 단위 IP 주소 프리픽스 (prefix) 범위에서 동일 버전을 다운로드하는 사용자들의 분포를 조사하였다. 이 결과 역시 효과적인 평판 시스템 설계에 도움을 줄 수 있는데, 특히 시빌 (Sybil) 공격[17]을 통해 평판 시스템을 무력화하려는 파일 오염 공격자들을 인지하는데 활용될 수 있다.

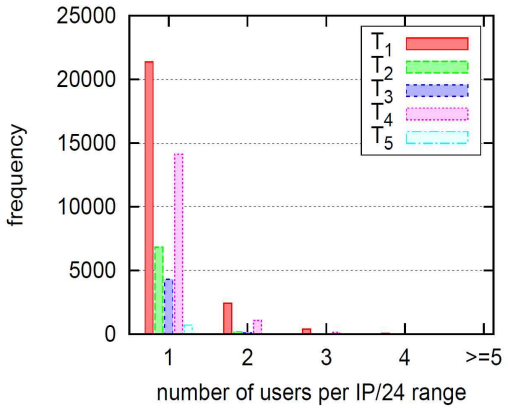

그림 6. 단위 IP 주소 프리픽스별 다운로더들 분포
Fig. 6. The number of users per IP/24 (Kad)

그림 6은 동일한 IP/24 프리픽스 범위에서 실험에 사용된 각 음악파일들의 최고 인기버전을 다운로드하는 사용자 수를 표시하고 있는데, 결과에서 보듯이 평균 1.1명에 불과했다. 결과를 얻기 위한 측정 기간이 각 콘텐츠 키의 재출판 주기인 5시간보다 두 배 긴 10시간이었음을 감안한다면 실제 그 수는 훨씬 적을 것으로 판단된다. 그림 5의 결과에서 보듯이 현재 Kad 네트워크에서 사설 IP 주소 사용자가 전체 사용자의 50%

에 육박하고 있음을 감안할 때 동일한 버전을 다운로드하는 사용자가 특정 IP 프리픽스 범위에 집중되는 경우가 거의 없다는 사실은 놀랄 만한 결과이다. 이 결과에 비추어 볼 때 만일 한 IP 프리픽스 범위에서 다수의 사용자가 동일한 버전을 출판하고 있다면 시빌 공격에 의한 파일 오염을 의심해 볼 수 있을 것이다.

2. 파일 오염 정도에 대한 측정결과 분석

다음으로 본 논문은 보완된 측정노드들로 하여금 키워드 출판 메시지에 대한 검증을 실시하게 하는 방식으로 Kad 네트워크에서의 파일 오염 정도를 분석하였다. 물론 콘텐츠 출판 메시지에 대해서도 동일한 실험을 하였으나, 측정 결과 현재 Kad 네트워크에서는 콘텐츠 출판을 통한 인덱스 오염공격보다는 키워드 출판을 통한 오염공격이 주를 이루고 있어, 본 논문에서는 키워드 출판 메시지에 대한 분석 결과만 언급한다.

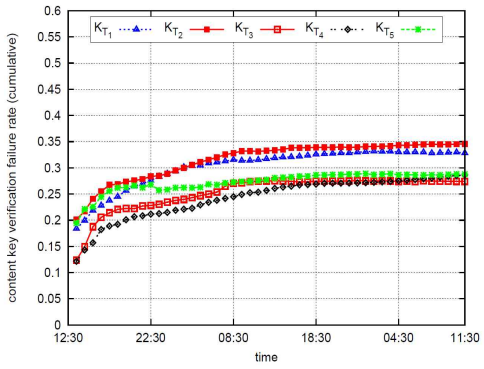

그림 7. 콘텐츠 검색 누적 실패율
Fig. 7. Cumulative content key verification failure rate

그림 7은 측정노드가 측정에 사용된 5개의 음악파일들에 대한 키워드 출판 메시지를 받을 때마다 출판 메시지에 담겨진 콘텐츠 키로 콘텐츠 탐색을 실시했을 때의 누적 실패율을 나타낸다. 결과에서 보듯이 측정노드가 콘텐츠 탐색 시 출판자 정보를 찾을 수 없는 콘텐츠 키의 비율이 적게는 27%에서 많게는 35%에 이르고 있음을 알 수 있다. 위 결과를 놓고 볼 때, 현재 Kad 네트워크의 인덱스 오염이 심각한 수준에 있음을 알 수 있다[3]. 이 결과를 통해 인덱스 노드들 (키워드 키 소유자 및 콘텐츠 키 소유자)로 하여금 수신된 출판 메시지의 내용을 검증하도록 함으로써 파일 오염 공격을 효과적으로 줄일 수 있으며, 나아가 개별적인 다운로더들이 원하는 파일을 찾는데 드는 시간과 노력을 줄일 수 있음을 알 수 있다. 따라서 향후 효과적인 파일 오염방지 시스템을 구현하기 위해서는 '인덱스 노드들이 저장하고 있는 출판자 및 파일에 대한 정보의 정확성을 어떻게 높일 수 있는가?'에 대한 연구가 필수적이라 하겠다.

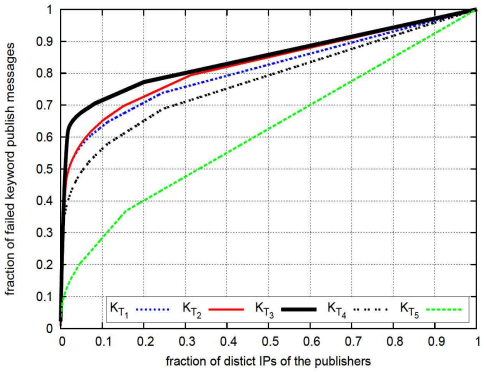

그림 8. 파일 오염 공격자들의 개별 IP 주소 비율 대비 가짜 키워드 출판 메시지 비율의 누적분포함수
Fig. 8. CDF of fraction of index polluters

그림 8은 가짜 키워드 출판 메시지 보내는 파일 오염 공격자들의 개별 IP 주소 비율 대비 가짜 키워드 출판 메시지 비율의 누적분포함수 (Cumulative Distribution Function, CDF)를 나타낸다. 이 그림에서 IP 주소들은 가짜 키워드 출판 메시지를 보내는 수에 의해 정렬되었다. 즉 가장 많은 가짜 키워드 출판 메시지를 보내는 IP 주소가 맨 앞에 위치한다. 결과에서 보듯이 4개의 유명한 mp3 음악파일들에 대해 가짜 키워드 출판 메시지의 60% 이상이 전체 IP 주소의 20% 내에서 출판되고 있음을 알 수 있다. 이는 4개의 유명한 음악파일에 대해 의도적으로 다수의 가짜 키워드 출판 메시지를 보내는 소수의 파일 오염 공격자들이 존재하고 있음을 명백하게 보여준다. 본 실험을 통해 4개의 유명한 음악파일에 대해 지속적으로 대량의 가짜 키워드 출판 메시지를 보내는 다수의 오염 공격자들의 IP 주소가 포함되어 있는 2개의 IP/24 프리픽스 범위가 발견되었는데, 이는 저작권 보호를 위해 음원 공급자에게 고용된 전문 파일 오염회사[1,2,12]로 추정된다. 재미있게도 상대적으로 유명하지 않은 5번째 음악 파일에서는 이러한 현상이 발견되지 않았다.

4) Ⅲ-3장에서 설명된 바와 같이, 파일 오염 공격자들은 출판 메시지에 잘못된 정보를 넣음으로써 다운로더들로 하여금 원하는 파일을 찾지 못하도록 유도한다.

VI. 관련 연구

파일 오염 공격에 대한 문제는 뉴욕 폴리텍 대학에서 2005년에 실시한 KaZaA 네트워크 성격과 파일 오염도 측정 연구에서 시작되었다[2]. 논문의 저자들은 저작권 보호를 위해 레코드 회사들에 의해 고용된 전문 파일 오염 회사들이 KaZaA 네트워크에서 다량의 오염된 파일들을 공유하고 있음을 발견하였다. 이 전문 파일 오염회사는 다운로더들로 하여금 오염된 파일들을 다운로드하도록 유도함으로써 KaZaA 시스템의 붕괴를 유도하고 있었다. 또한 당시의 파일 오염 방법은 주로 콘텐츠 오염 (Ⅲ-1장)과 메타데이터 오염 (Ⅲ-2장)이 주를 이루고 있음을 보였다. 2006년에 이르러 파일 오염 공격의 패턴은 인덱스 오염 (Ⅲ-3장)이라는 새로운 방법으로 진화해 파일 오염 공격자의 부담은 줄어들고, 그 효과는 늘어나게 되었다. 이와 관련해 폴리텍 대학 연구진은 구조화되지 않은 파일 공유 시스템인 FastTrack과 DHT 기반 시스템인 Overnet에서의 오염 문제를 연구하여 인덱스 오염이 새로운 문제로 대두되고 있음을 보였다[3]. 본 논문은 이 두 연구와 그 맥락을 같이하고 있으나 그 연구 대상이 Kad 네트워크라는 점이 가장 큰 차이점이라 하겠다.

파일 오염 방지의 중요성 때문에 현재 다양한 방면에서 활발한 연구가 진행되고 있다. 이러한 연구들은 오염된 파일의 확산에 대한 이해를 돕는 파일 오염 역학 모델링 연구[12,18,19], 출판자에 대한 평판을 바탕으로 다운로드할 파일을 결정하는 피어 (peer) 평판 접근방법[16,20], 특정 파일 버전에 대한 평판을 바탕으로 다운로드할 파일을 결정하는 파일 버전 평판 접근방법[21,22,23], 그리고 피어와 파일 버전 평판이 결합된 형태의 혼성 접근방법[22,24] 등으로 구분될 수 있다. 이러한 연구들과는 달리, 본 논문은 DHT 기반의 파일 공유 P2P 시스템에서 사용자 행동양식 및 파일 오염에 대한 실질적인 측정 연구에 중점을 두고 있다.

VII. 결론 및 향후 연구방향

본 논문은 Kad 네트워크에 대한 측정연구를 통해 DHT 기반의 파일 공유 P2P 시스템 사용자의 행동양식을 분석하고, 시스템 내 파일 오염 수준을 인덱스 오염을 중심으로 측정하였다. 이번 연구를 통해 다음과 같은 Kad 네트워크 이용자들의 행동양식을 이해할 수 있었다. 먼저 Kad 사용자들의 경우 주로 낮에 파일을 주고받으며, 새로운 파일의 공유보다는 이미 존재하는 파일의 확산이 지배적이었다. 버전 인기도가 지프 분포 (Zipf distribution)를 따른다는 사실을 통해 Kad 사용자들이 버전 인기도에 의존해 다운로드할 파일을 결정하고 있음을 유추할 수 있었고, 전체 사용자의 약 50%에 달하는 사용자들은 NAT 혹은 방화벽 뒤에 위치하고 있음을 확인할 수 있었다. 또한 많은 사용자들이 사설 IP 주소를 사용하고 있음에도 불구하고 좁은 IP 주소 프리픽스 (IP/24) 범위에서 같은 버전의 파일을 다운로드하는 사용자의 수는 평균 1.1명 이하로 매우 적었다. 다음으로 Kad 네트워크에 대한 인덱스 오염 분석 결과를 통해 인덱스 노드에 지속적으로 미끼 (decoy) 정보를 삽입하는 파일 오염 공격자들이 존재함을 확인할 수 있었는데, 이런 파일 오염 공격자들의 경우 전문 오염회사로 추정된다. 인덱스 오염의 결과로 Kad 네트워크 안에서는 약 35%에 육박하는 인덱스 정보가 이미 오염되어 있으며, 그 결과 Kad 네트워크 사용자들은 원하는 파일을 찾기 위해 상당한 시간과 노력을 낭비하고 있음을 알 수 있었다.

이와 같은 측정 연구 성과를 바탕으로, 앞으로 우리는 DHT 기반 파일 공유 P2P 시스템에서 인덱스 노드가 저장하고 있는 파일 및 출판자 정보를 정화하는 방식의, 새로운 파일 오염 방지기법 개발에 관한 연구를 진행할 계획이다.

참고문헌

[1] N. Christin, A. S.Weigend, and J. Chuang, "Content vailability, pollution and poisoning in file sharing peer-to-peer networks," in ACM EC'05, pp. 68-77, 2005.

[2] J. Liang, R. Kumar, Y. Xi, and K. W. Ross, "Pollution in p2p file sharing systems," in IEEE INFOCOM'05, Miami, FL, 2005.

[3] J. Liang, N. Naoumov, and K.W. Ross, "The index poisoning attack in p2p file sharing systems," in IEEE INFOCOM'06, 2006.

[4] emule morph, *http://emulemorph.sourceforge.net*

[5] R. Brunner, "A performance evaluation of the kad-protocol," Master's thesis, University of Mannheim, Sophia-Antipolis, France, 2006.

[6] eMule Project, *http://www.emule-project.net.*

[7] aMule Forum, *http://forum.amule.org/.*

[8] P. Maymounkov, and D. MaziItes, "Kademlia: A peer-to-peer information system based on the xor metric", in IPTPS'02, pp. 53-65, 2002

[9] D. Stutzbach, and R. Rejaie, "Improving lookup

performance over a widely deployed dht," in IEEE INFOCOM'06, pp. 1-12, 2006.
[10] M. Steiner, T. En-Najjary, and E. W. Biersack, "Exploiting kad: Possible uses and misuses," in ACM SIGCOMM Computer Communication Review 37, 65-70, 2007
[11] P. Dhungel, X. Hei, K. W. Ross, and N. Saxena, "The pollution attack in p2p live video streaming: measurement results and defenses," in the workshop on Peer-to-Peer streaming and IP-TV, Japan, 2007.
[12] U. Lee, M. Choi, J. Cho, M. Y. Sanadidi, and M. Gerla, "Understanding pollution dynamics in p2p file sharing," in IPTPS'06, Santa Babara, USA, 2006.
[13] MD4, *http://en.wikipedia.org/wiki/MD4.*
[14] Top 10 songs, *http://top10songs.com/* (June 2008).
[15] K. Shin, D. S. Reeves, I. Rhee, and Y. Song, "Winnowing : Protecting p2p systems against pollution by cooperative index filtering," Tech. Rep. TR-2009-2, North Carolina State University, 2009.
[16] C. Costa, V. Soares, J. Almeida, and V. Almeida, "Fighting pollution dissemination in peer-to-peer networks," in ACM SAC'07, Seoul, Korea, pp. 1586-1590, 2007.
[17] J. R. Douceur, "The sybil attack," in IPTPS'02, Cambridge, MA, 2002.
[18] D. Dumitriu, E. Knightly, A. Kuzmanovic, I. Stoica, and W. Zwaenepoel, "Denial-of-service resilience in peer-to-peer file sharing systems," in ACM SIGMETRICS'05, Ban, Alberta, Canada, pp. 38-49, 2005.
[19] R. Kumar, D. D. Yao, A. Bagchi, K. W. Ross, and D. Rubenstein, "Fluid modeling of pollution proliferation in p2p networks," ACM SIGMETRICS Performance Evaluation Review, 335-346, 2006.
[20] S. D. Kamvar, M. T. Schlosser, and H. Garcia-molina, "The eigentrust algorithm for reputation management in p2p networks", in the Twelfth International World Wide Web Conference, pp. 640-651, 2003.
[21] K. Walsh, and E. G. Sirer, "Experience with an object reputation system for peer-to-peer file sharing," in USENIX NSDI'06, San Jose, CA, 2006.
[22] C. Costa, and J. Almeida, "Reputation systems for fighting pollution in peer-to-peer file sharing systems," in IEEE P2P'07, 2007.
[23] E. Zhai, R. Chen, Z. Cai, L. Zhang, E. K. Lua, H. Sun, S. Qing, L. Tang, and Z. Chen, "Sorcery: Could we make p2p content sharing systems robust to deceivers?", in the 9th International Conference on Peer-to-Peer Computing, 2009.
[24] N. Curtis, R. Safavi-Naini, and W. Susilo, "X2rep : Enhanced trust semantics for the xrep protocol", in ACNS'04, 2004.

저 자 소 개

신 규 용

1996년 3월 육군사관학교 전산학 학사.
2000년 2월 한국과학기술원 전산학 석사 (ATM 네트워크).
2009년 12월 노스캐롤라이나 주립대학 전산학 박사 (분산 시스템 보안).
현재 : 육군사관학교 전자정보학과 정보과학 조교수
관심분야 : 컴퓨터 네트워크, 분산 시스템, 인센티브, 네트워크 보안
Email : yessss@gmail.com

유 진 철

1989년 3월 육군사관학교 전산학 학사.
1993년 7월 아이오와 주립대 통계학 석사 (게임이론).
2003년 5월 펜실베니아주립대 컴퓨터공학 박사 (고성능/저전력 시스템)
현재 : 육군사관학교 전자정보학과 정보과학 부교수
관심분야 : 고성능 컴퓨팅, 저전력 시스템, 정보 보안
Email : jyoo@kma.ac.kr

이 종 덕

2005년 3월 육군사관학교 전산학 학사.
2009년 5월 버지니아 주립대 컴퓨터과학 석사 (무선 센서 네트워크 전공).
현재 : 육군사관학교 전자정보학과 강사
관심분야 : 무선 센서 네트워크, 보안
Email : jdlee@kma.ac.kr