빅데이터 분석기법을 활용한 연계계좌 데이터 구축에 관한 연구*

양 철 원 (단국대학교 부교수)**

〈요약〉

최근 불공정거래 사건의 가장 큰 특징은 혐의계좌 수의 증가이다. 불공정거래 조사 건수는 매년 감소하고 있음에도 평균 혐의 계좌수는 계속 증가하고 있다. 이는 많은 불공정거래 행위가 단독 계좌보다는 복수의 연계계좌를 통해 이루어지고 있음을 의미이다. 시장감시 업무도 단일계좌보다는 연계계좌를 찾는 것이 점점 중요해지고 있다. 이에 맞추어 본 연구에서 3단계에 걸친 연계계좌 적출방법을 제안하였다. 첫 단계는 2계좌 간의 연계성을 계산하기 위해 주식 거래 및 호가자료를 사용하여 20개의 연계지표를 개발하였다. 두 번째 단계는 연계지표를 결합하거나 특정 사건일을 사용하여 계좌쌍을 필터링함으로써 혐의 연계계좌군을 특정하였다. 마지막 단계는 특정된 연계계좌들에 대한 사회네트워크 분석(social network analysis) 통해 계좌 간의 연결구조 및 연결강도, 그리고 핵심 연계계좌군을 파악하였다. 연계계좌에 대한 분석은 많은 계산량과 시간을 요구하기 때문에 시장감시기구는 연계계좌들을 미리 인지해 놓고 있다가 거래상 이상징후 발생할 시에 이를 중심으로 조사할 수 있다. 또한 실제 불공정거래사건 처리 후 연계계좌를 업데이트함을 통해 더 정교한 연계계좌 데이터베이스를 구축할 수 있을 것이다. 이런 과정을 통해 불공정거래에 대한 시장감시 기능을 더 강화할 수 있을 것으로 기대된다.

핵심 단어 : 불공정거래, 연계계좌, 연계지표, 사회네트워크 분석, 시장감시

JEL 분류기호: G14, G18, G28

* 본 연구는 2018년도 한국거래소-한국증권학회 건전증시 연구지원사업의 지원을 받아 수행되었습니다. 본 논문에 대하여 유익한 조언을 해 주신 송수영, 배경훈 교수님, IBK투자증권 신호철 상무님, 한국거래소 김현철 상무님과 2018년 12월 한국거래소 건전증시포럼 참가자들과 감사드립니다.

** 연락담당 저자. 주소: 경기도 용인시 수지구 죽전로 152 단국대학교 경영학부, 16890;
E-mail: yang@dankook.ac.kr; Tel: 031-8005-3437; Fax: 031-8021-7208.

투고일 2019-01-04; 게재확정일 2019-01-29

1. 서론

자본시장의 기능은 높은 수익성과 성장성이 기대되는 기업이 낮은 비용으로 자본을 조달할 수 있게 함으로써 사회의 생산효율을 극대화하고 부를 창출하는 것이다. 하지만, 자본시장의 불공정거래는 시장의 가격발견 기능을 훼손하며 그 결과 비효율적인 자원배분을 가져오기 때문에 그 폐해가 심각하다. 금융감독원은 보도자료를 통해 2017년 중 총 137건의 불공정거래 사건에 대한 조사를 종결하여 그 중 77건을 검찰에 고발, 통보하였다고 보고하면서, 최근 불공정거래 조사사건의 가장 큰 특징으로 혐의계좌 수 증가를 지적하고 있다.¹⁾ 불공정거래 조치 건수는 2015년 195건, 2016년 172건, 2017년 137건으로 매년 감소하고 있지만, 1건당 평균 혐의 계좌수(거래소 통보 사건)는 2015년 18개, 2016년 22개, 2017년 30개로 계속 증가하고 있다. 이는 불공정거래 행위가 단독 계좌 보다는 복수의 계좌를 통해 이루어지고 있음을 의미한다. 즉, 단일 범행보다는 고도로 조직된 팀, 일명 ‘작전’을 수행하는 그룹들이 증가하는 것이다.

이에 대해서는 몇 가지 이유를 생각해 볼 수 있다. 첫째, 복수의 계좌를 사용하면 불공정행위를 찾아내는 것이 쉽지 않다. 즉, 적발 가능성이 낮아지는 것이다. 보통 시장감시시스템은 가격이나 주문 등에 이상징후가 나타날 때 그 원인을 제공한 계좌를 추적하는 방식으로 이루어진다. 즉, 가격충격이 클수록 감시대상이 될 가능성이 높다. 불공정거래자는 한 계좌를 통해 대규모 거래를 하는 것보다 여러 계좌로 주문을 쪼개서 제출하는 것이 가격충격을 완화시킬 수 있다. Barclay and Werner(1993)도 정보를 가진 거래자가 자신의 거래에서의 가격충격을 최소화하기 위해서 대규모 주문보다는 중규모의 주문으로 분할 제출함으로써 자신을 숨기려는 경향이 있음을 보고하고 있다. 이를 은닉거래 가설이라고 한다. 둘째, 여러 사람이 모일 경우 자금 동원력이 더 커지게 된다. 셋째, 미디어의 발달 등으로 인해 여러 사람이 참여하더라도 빠른 연락이 가능하며 효과적인 작전을 수행할 수 있다.

이런 현상들을 볼 때, 시장감시 업무에 있어서도 기존의 단일계좌 중심보다도 연계계좌 중심의 조사업무의 중요성이 커지고 있다. 하지만 연계계좌 조사는 단일계좌보다 그 계산량이 기하급수적으로 늘어나게 된다. 주식시장에서만 해도 단일계좌가 수백만 개에 이르는데 이들의 짝을 조사한다고 하는 것은 엄청난 계산량을 요구하게 된다. 따라서, 어떤 사건이 발생했을 때 연계계좌를 조사하고 하는 것은 시간도 많이 걸리고 분석 결과도 시원치 않을 때가 많다. 본 연구는 이를 돕고자 3단계에 걸친 연계계좌 적출방법을 제안하였다.

본 연구에서 제시하는 구체적인 연계계좌 적출방법은 다음과 같다. 연계계좌를 통한 불공정 행위는 다양한 조합으로 발생할 수 있다. 2계좌끼리 연계한 불공정 행위도 가능하며, 3계좌, 4계좌, 결국 n계좌까지 연계가 가능하다. 모든 조합을 검증하기보다는 가장 기초가 되는 2계좌 사이의 연계에 대한 분석을 시작점으로 하였다. 먼저 2계좌 간의 연계성을 측정해 주는 다양한 연계지표를 개발하였다. 본 연구는 한국증권거래소에서 제공한 거래자료와 호가자료를 사용하여 20개의 연계지표를 계산하였다. 예를 들면, 서로 같은 종목을, 같은 날에 많이 거래하는 계좌들은

1) 금융감독원 보도자료, 2018. 2. 5, 2017년 자본시장 불공정거래 주요 적발사례 및 투자 유의사항.

다른 계좌에 비해 연계성이 높다. 이런 계좌들이 불공정거래 중 하나인 통정매매나 가장매매에 관여할 가능성도 더 클 것이다. 이와 비슷한 개념으로 다양한 연계성을 측정할 수 있는 20개의 연계지표를 계좌쌍들에 대해 계산하였다.

두 번째 단계는 앞 단계에서 구한 계좌쌍들을 필터링하여 혐의 연계계좌군을 특정하는 것이다. 여기에는 계좌들 사이의 2개 이상의 연계지표를 결합하여 필터링하는 방법과 특정 사건일을 이용하는 방법을 사용하여 연계계좌군을 특정하였다. 위의 필터링을 적절하게 적용하면 특정 불공정거래행위에 대해 혐의가 있는 30~100개 정도의 연계계좌들로 조사 범위를 좁힐 수 있다.

마지막 단계는 특정된 연계계좌들을 대상으로 사회네트워크 분석(social network analysis)을 통해 계좌들이 서로 어떻게 연계되어 있는지 분석하였다. 이를 통해 계좌 간의 연결구조 및 연결강도 등을 측정할 수 있으며, 어떤 계좌가 네트워크의 중심인지 그 중심성도 찾아낼 수 있다. 또한 컴퍼넌트 분석을 통해 네트워크 안에 존재하는 핵심 연계계좌군에 대해 파악할 수 있다.

이를 기반으로 실제 시장감시업무를 진행한다면 업무효율성을 높일 수 있을 것이라 기대된다. 특히 사회네트워크 분석과 같은 빅데이터 분석은 불공정거래 연계계좌를 인지하는데 적합하다. 증권분야만큼 대용량의 데이터를 매일 생성하는 분야도 없다. 또한 빅데이터 분석을 통해 연계계좌들을 미리 인지하고 데이터베이스화할 수 있다. 연계계좌에 대한 분석은 많은 계산량과 시간을 요구하기 때문에 기존 거래를 통해 연계계좌를 미리 인지해 놓을 수 있다면, 실제 이상징후 발생 시 구축해 놓은 연계계좌를 중심으로 조사할 수 있다. 예를 들면, 경찰도 조직폭력배 조직과 조직원에 대한 데이터베이스를 미리 구축하고 있으며, 조직폭력 사건 발생 시 기존 자료를 중심으로 관련자나 배후세력을 수사한다. 연계계좌에 대한 적출과 데이터베이스화를 통해 시장감시업무에 있어서도 이런 접근이 가능해질 것이다. 또한 불공정사건 발생시 연계계좌에 대한 업데이트를 통해 더 정교한 정보를 확보할 수 있다. 불공정거래 사건을 통해 밝혀진 실제 연계계좌들을 사용함으로써 연계계좌 데이터베이스를 검증할 수 있으며 더 발전시킬 수 있다.

본 연구는 다음과 같은 면에서 기여할 수 있으리라 사료된다. 첫째, 실제 업무에서 연계계좌를 통한 불공정거래 행위를 적발하는데 기여할 수 있다. 최근 불공정거래가 단독 계좌 보다는 복수의 계좌를 통해 이루어지고 있음을 생각할 때 연계계좌를 적출하고 이를 데이터베이스화하는 것이 실제 조사업무에 큰 도움이 될 것이다. 일반 범죄 분야에서도 대포폰의 적출, 테러리스트 적출 등의 범죄 분야에서 빅데이터 분석을 통해 성과를 올리고 있는 현실을 생각할 때, 가장 지능적이라 할 수 있는 증권 범죄에도 이런 기법들이 도입되어야 한다고 생각한다.

둘째, 시장에서 불공정거래행위 예방 효과를 가져 올 수 있다. 시장감시기구가 연계계좌들의 데이터베이스화해서 가지고 있다는 사실은 시장참여자들의 불공정행위를 억제하는 효과를 줄 수 있다. 또한 연계계좌의 정확성은 연계계좌에 대한 지속적인 업데이트를 통해 더 향상시킬 수 있다. 이런 광범위하고 정교한 정보를 시장감시기구가 가지고 있다는 사실이 시장참여자들에게 알려진다면, 일종의 신호효과를 주어 잠재적 불공정행위를 예방하는 효과를 줄 수 있다.

셋째, 빅데이터 분석기법을 다른 금융분야에 적용할 수 있다. 자본시장의 증권거래는 참여자나 거래량이 많기 때문에 불공정거래 혐의자를 찾아내는 것이 쉬운 일은 아니다. 빅데이터를 사용하여

증권거래에서 연계계좌를 적출하여 성과를 낼 수 있다면, 이 보다 자료 크기가 훨씬 작은 은행이나 보험 등의 다른 분야에도 적용가능하리라 사료된다.

본 논문은 다음과 같이 구성된다. 제 2장은 문헌연구와 연계계좌 적출을 위한 연구설계를 소개한다. 제 3장에서는 연계지표와 사회네트워크 분석 방법론 등에 대해서 구체적으로 설명한다. 제 4장은 자료를, 제 5장에서는 분석결과를 소개한다. 제 6장에서는 연구를 마무리한다.

2. 문헌연구 및 연구설계

2.1 문헌연구

불공정거래의 유형으로는 시세조종, 미공개정보이용(내부자 거래), 부정거래 행위, 단기매매 차익거래, 주식소유 및 대량보유 보고의무 위반, 신고·공시 의무 위반 등 다양하다. 하지만, 일반적으로 ‘불공정거래’라 하면 시세조종과 미공개정보이용 두 가지를 지칭하며, 학술연구들도 주로 시세조종과 미공개정보이용(내부자 거래)을 다루고 있다. 시세조종은 정보를 이용하기 보다는 주가조작자들의 수급에 의해서 시세를 왜곡하는 행위인 반면, 미공개정보이용은 정보를 이용하여 불법적인 이익을 취득하는 행위로 정의된다. 예를 들면, 기업 공시 전에 미리 그 정보를 이용하여 거래하는 행위가 해당된다.

시세조종에서는 주가를 상승시키는 행위가 주가를 하락시키는 경우보다 일반적이다. 주가를 하락시키기 위해서는 가지고 있는 주식을 대량 매도하든지 공매도를 해야 하는데, 이는 주식을 매수하는 것보다 더 많은 비용을 요구하기 때문이다. 수급에 의한 시세조종으로는 주식을 매집함으로 주가를 끌어올리고 이를 매도하는 행위(일명 pump and dump)가 가장 대표적이다. Khwaja and Mian(2005)은 파키스탄 주식시장에서 증권사들이 다른 투자자들보다 연 50~90% 더 높은 수익률을 얻는데, 이 수익률의 절반 정도가 pump and dump 기법의 주가조작 행위로 설명될 수 있음을 보여주었다. 수급에 의해서가 아닌 정보를 이용한 주가조작도 가능하다. 예를 들면, 주가조작자들이 방송이나 신문기사 등을 이용한 허위정보 유포하여 주식가격을 조작할 수 있다.

시세조종에 대한 문헌들은 특히 종가시세 조종과 관련된 연구들이 많다. 종가시세 조종에 관한 대부분의 연구는 펀드매니저가 자신의 성과 상승을 위해 인위적으로 종가를 상승시키고자 하는 유인이 존재함을 지지한다. Carhart et al.(2002)은 미국의 펀드와 주식 일종거래자료를 사용하여 분기 말의 펀드순자산 가치가 비정상적으로 크며, 펀드매니저들이 분기 말일 종가 전 30분 동안 보유 주식을 아주 활발하게 거래함을 보여주었다. 그들은 이런 현상이 펀드 매니저들이 분기 말 평가에 대비해 포트폴리오의 가치를 인위적으로 증가시키려는 주가조작 때문임을 주장하였다. Bernhardt and Davies(2005)도 분기 마지막 날의 수익률이 다른 평일보다 유의미하게 높으며 이런 현상이 펀드가 보유한 종목일수록 더 강함을 발견하였다. 이를 펀드 매니저들이 종가시세 조종을 하는 증거로 해석하였다. Comerton-Forde and Putnins(2011)는 미국, 캐나다의 감독기관이 종가시세 조종으로 기소한 사례에 해당하는 31개 종목 총 184건을 분석하였다. 분석결과, 시세조종행위가 수익률, 수익률의 반전현상, 거래빈도, 호가스프레드 등을 유의미하게 증가시킴을 발견하였다. 이와 비슷하게 박연길 외 2인(2013)은 2011년 한국의

법원에서 펀드매니저의 종가시세 조종에 대해 유죄를 선고한 사건의 14개 종목을 분석하였다. 그들은 종가시세 조종이 주가와 거래량을 유의미하게 상승시키며 다음날에는 수익률이 반전됨을 보여주었다. 양철원, 유지연(2016)은 한국의 선물과 옵션만기일의 만기일 효과에 지수차익거래가 유의미한 역할을 하고 있음을 보여주었다. 또한 이들은 만기일 코스피200 종가결정 시간대의 주문들을 분석한 결과, 연계시세조종을 의심할만한 주문들이 평일에 비해 유의적으로 증가함을 발견하였다. 양철원, 유지연(2017)은 2010년 옵션만기일 발생한 도이치은행의 대규모 주식매도로 인한 ‘옵션쇼크 사건’을 연계시세조종의 관점에서 분석한 사례연구를 제시하였다. 양철원(2018)은 파생상품 만기일의 현물 주가조작에 대한 Kumar and Seppi(1992) 모형의 예측을 한국 시장의 자료를 사용하여 검증하였다.

종가시세 조종에 대한 이론 모형들도 존재한다. Hillion and Suominen(2004)는 대리인 이론에 기반하여 브로커가 고객들이 자신의 능력을 과대평가하도록 만들기 위해 종가 시세를 조종할 유인이 발생할 수 있음을 모형화하였다. Kumar and Seppi(1992)는 정보불균형 하의 현금결제 방식의 선물시장을 모형화하여 주가조작가 현물시장의 종가시세를 조종을 통해 미리 보유하고 있던 선물포지션에서 이익을 취하는 전략이 균형상태임을 증명하였다. 이는 현물과 선물을 동시에 보유한 투자자가 파생상품에서 이익을 얻기 위해서 현물의 종가를 조종하려는 유인이 존재함을 보여준다.

한편 빅데이터 관련 기술을 고찰한 연구들이 존재한다. 김정숙(2012)은 국내외 빅데이터 활용 현황과 함께 빅데이터 분석기법과 인프라 기술 등을 살펴보고 있다. 김재생(2014)은 빅데이터 분석기술을 오픈소스 분야, 상용솔루션 분야, 클라우드 서비스 분야로 나누어서 살펴보고, 실제 활용사례들을 제시하고 있다. 김상락, 강만모(2014)는 빅데이터 분석기술과 솔루션을 살펴보고 실제 활용 사례들을 제시하고 있다. 김석수(2014)는 빅데이터 활용 기술로 데이터 저장 레이어, 데이터 처리 레이어, 데이터 분석 레이어, 시각화 레이어 등의 각 단계를 Hadoop을 통하여 데이터를 처리한 후 처리된 데이터를 Mahout으로 실행하였다. 실제로 금융 상품에 가입된 고객의 여러 특성을 이용하여 군집분석을 시행하고 각 고객에 따른 금융 상품을 추천할 수 있는 사례를 개발하여 시각화하여 보여주고 있다.

2.2 연구설계

본 연구는 불공정거래 행위가 단일 주가조작자에게만 아니라 다수가 공모하여 일어날 수 있음을 가정하여 연계계좌를 적출하는 방법을 제안하고자 한다. 결국 혐의계좌를 적출하는 방법은 수많은 계좌들 중에서 혐의없는 계좌들을 제거하고 가장 가능성 있는 소수의 계좌들을 남기는 필터링 과정의 연속이라고 할 수 있다. 본 연구에서 한국 주식시장의 3개월간의 자료를 분석했을 때 활동한 주식계좌의 수는 약 3백만 개에 이른다. 이런 많은 계좌 간의 연계성을 파악하여 연계계좌를 적출하는 것은 한 번의 필터링으로는 어렵다. 따라서 본 연구에서는 <그림 1>과 같이 3단계에 걸쳐서 연계계좌를 적출하는 방법을 제안하였다.

첫 번째 단계는 계좌쌍별로 단일 연계지표를 계산한다. 일차적으로 모든 거래일을 대상으로 하여 두 계좌 간의 연계지표를 만들어서 계좌쌍 간의 관계에 대해서 측정한다. 본 연구에서는

주가조작 공모자들의 거래나 호가 행태에서 고안한 20개의 연계지표를 계산하고 이들의 분포를 통해 사용 가능여부를 판단해 보았다.

두 번째 단계는 단일지표의 결합 또는 사건일을 사용하여 계좌쌍을 줄이는 방법으로 연계 계좌들을 특정하고자 하였다. 두 개 이상의 단일지표를 결합함으로써 연계 혐의가 큰 소수의 계좌쌍을 필터링할 수 있다. 또는 상한가일, 이익공시일 등의 사건일을 특정하여 혐의 계좌들을 추출함으로써 계좌의 수를 줄일 수 있다. 이는 모든 거래일을 대상으로 분석하기보다는 주가조작 가능성이 높은 거래일만을 사용하는 것이 더 효과적일 수 있기 때문이다. 시세조종의 가능성이 높은 거래일로는 상한가가 형성된 날, 급격한 수익률의 변동이 발생한 날, 상당히 오랜 기간 주가가 연속적으로 상승하거나 하락하는 경우 등을 들 수 있다. 미공개정보이용의 가능성이 높은 거래일로는 이익공시일 전 거래일, 인수, 합병 등 중요한 정보공시일 전 거래일, 기업에 대한 신문기사나 방송이 나온 날 주변 거래일 등을 생각할 수 있다.

마지막 3단계는 빅데이터 분석기법 중 사회네트워크 분석을 적용하였다. 사회네트워크 분석을 통해 계좌들 간의 구체적인 관계 및 역학에 대하여 파악할 수 있으며, 특히 컴퍼넌트 분석을 통해 네트워크 안에 존재하는 더 세부적인 연결망을 특정할 수 있다.

<그림 1> 연계시세조종 혐의계좌 적출 방법

3. 연구방법론

3.1 연계지표의 개발

여기서는 계좌쌍 연계지표에 대해 설명하고자 한다. 본 연구에서는 거래자료를 사용한 연계지표 10개, 호가자료를 사용한 연계지표 10개를 개발하였다. 여기에 소개한 20개 연계지표는 기본적으로 두 계좌와의 관계에 관한 것이다. 우리가 통계학에서 흔히 사용하는 공분산(covariance)나 상관관계수(correlation)와 비슷한 개념이라고 이해하면 된다. 다만 계산에 사용할 입력 변수로서 연계 불공정거래를 잡아낼 수 있는 다양한 계좌 활동 변수들을 고려하였다.

<표 1>은 거래자료를 이용한 10개 지표들에 대한 설명과 구체적인 계산방법을 제시하고 있다. 예를 들어, 첫 번째 연계지표인 ‘공통 종목_거래일(전체) 지표’를 살펴보자. 이는 두 계좌가 거래하는 종목과 거래일을 사용한 계좌간 연계성에 관한 변수이다.

$$\text{공통 종목_거래일 지표} = \frac{\text{공통 종목_거래일}}{\text{Min}(i\text{계좌의 종목_거래일}, j\text{계좌의 종목_거래일})} \quad (1)$$

공통 종목_거래일은 계좌 i와 계좌 j가 동일하게 거래한 종목 및 거래일의 수이다. 두 계좌가 동일한 2개의 종목을 한 거래일에 호가한 경우나 1개의 종목을 동일한 두 거래일에 호가한 경우 종목-거래일은 2로 계산한다. 두 계좌가 같은 종목을 같은 거래일 동일하게 거래할수록 계좌간의 연계성이 높다고 할 수 있다. 투자자들이 많은 종목을 매매할 경우, 연계 계좌가 아님에도 불구하고 종목지표가 높아진다는 단점이 있다. 따라서, 분모로 두 계좌의 종목 및 거래일 중에서 작은 수치를 사용하여 연계지표를 표준화하고자 하였다. 반면 이 연계지표는 광범위한 종목, 기간을 대상으로 측정하기 때문에 포괄적인 변수라는 장점이 있다. 실제 변수는 전체 거래자료를 사용한 전체지표, 시가와 종가 거래만을 사용한 시가, 종가지표 각각 3개를 계산하였다.

네 번째로 통정거래일 연계지표를 고안하였다. 통정거래는 주가조작자가 사전에 약속한 특정 상대방과 미리 짜고 주식을 매매 하여 거래가 활발한 것처럼 위장하는 행위이다. 통정 종목_거래일은 두 계좌가 특정 종목에 대해서 특정 거래일에 거의 동일한 금액을 한 계좌는 매수하고 한 계좌는 매도한 종목_거래일을 계수하였다. 다음과 같이 두 계좌의 해당 종목의 해당 거래일에 순매수 거래금액의 합이 거래량의 5% 이내이면 통정거래일로 인식하였다.

$$\text{통정 종목_거래일} = \text{다음을 만족하는 계좌 } ij \text{의 공통 종목_거래일의 수} \quad (2)$$

$$\text{if } \left| \frac{\text{계좌 } i \text{의 순매수금액} + \text{계좌 } j \text{의 순매수금액}}{\text{계좌 } i \text{의 거래금액} + \text{계좌 } j \text{의 거래금액}} \right| < 0.05$$

두 계좌가 동일 거래일에 동일 종목을 거래하면 통정매매의 가능성이 높아지는 것이다. 이런 거래가 여러 번 반복된다면 통정매매로 인한 연계계좌일 가능성은 더 높아질 것이다. 앞에서 소개한 ‘공통 종목_거래일’ 지표가 혐의계좌를 광범위하게 포착하는 것에 반해 ‘통정 종목_거래일’은 연계혐의를 매우 엄격하게 적용하는 것이다.

다음 3지표는 ‘공통 거래금액관여율(매수/매도/순매수)’ 연계지표이다. 다음의 식 (3)은 ‘공통 매수 거래금액관여율’에 대해 보여주고 있다.

$$\text{공통 매수거래금액관여율} = \sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 매수 거래금액관여율의 곱} \quad (3)$$

$$\text{계좌 } i \text{의 매수 거래 금액 관여율} = \frac{\text{계좌 } i \text{의 해당종목에 대한 매수거래금액}}{\text{해당종목의 총 매수거래금액}}$$

특정 계좌의 매수 거래금액관여율은 해당 종목의 총 매수거래금액 중에서 그 계좌가 매수한 거래금액의 비율로 정의한다. 두 계좌에 대해서 이 비율을 곱하여서 두 계좌의 해당 종목에 대한 매수비중이 클수록 더 큰 값을 가지도록 하였으며, 이 값을 가지는 공통거래일에 대해서 합산함으로써 더 많은 거래일에 걸쳐서 동일한 거래를 할수록 연계지표의 값이 더 커지도록 고안되었다. 매도 거래금액과 순매수 거래금액에 대해서도 동일한 방법으로 연계지표를 계산하였다. 연계지표의 값이 클수록 두 계좌의 연계성이 더 강함을 의미하게 된다.

마지막 ‘공통 거래집중도(전체/시가/종가)’ 연계지표는 앞의 지표들과 다르게 그 계좌 자체의 거래특성에 초점을 두고 있다. 다음의 식 (4)은 공통 거래집중도에 대해 보여주고 있다.

$$\text{공통 거래집중도} = \sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 거래집중도의 곱} \quad (4)$$

$$\text{계좌 } i \text{의 거래집중도} = \frac{\text{계좌 } i \text{의 해당종목의 당일거래금액}}{\text{계좌 } i \text{의 당일거래금액}}$$

먼저 특정 계좌가 그 종목에 얼마나 집중하여 거래하고 있는지를 측정한 후, 이를 두 계좌 사이에 곱하여서 연계지표를 산출하였다. 수급에 의한 주가조작을 하려고 할 경우, 주가조작자가 여러 종목을 거래하기 보다는 한 종목에 집중하여 하는 특성을 고려하여 위의 연계지표를 고안하였다. 연계지표의 값이 클수록 두 계좌가 해당 종목에 집중하는 연계성이 더 강함을 의미한다.

〈표 1〉 거래자료를 이용한 계좌쌍 연계지표

번호	연계지표	설명
1	공통 종목_거래일(전체)	공통 종목 거래일 $Mn(i\text{계좌의 종목 거래일}, j\text{계좌의 종목 거래일})$
2	공통 종목_거래일(시가)	공통 종목 시가거래일 $Mn(i\text{계좌의 종목 시가거래일}, j\text{계좌의 종목 시가거래일})$
3	공통 종목_거래일(종가)	공통 종목 종가거래일 $Mn(i\text{계좌의 종목 종가거래일}, j\text{계좌의 종목 종가거래일})$
4	통정 종목_거래일	계좌 i, j 의 종목 순매수거래량 합이 거래량의 5%이내인 종목_거래일의 수
5	공통 거래금액관여율(매수)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 매수 거래금액관여율의 곱}$ 매수거래금액 관여율 = $\frac{\text{계좌 } i \text{의 해당종목에 대한매수거래금액}}{\text{해당종목의총매수거래금액}}$
6	공통 거래금액관여율(매도)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 매도 거래금액관여율의 곱}$
7	공통 거래금액관여율(순매도)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 순매수거래금액관여율의 곱}$
8	공통 거래집중도(전체)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 거래집중도의 곱}$ 계좌 i 의 거래집중도 = $\frac{\text{계좌 } i \text{의 해당종목의 당일거래금액}}{\text{계좌 } i \text{의 당일거래금액}}$
9	공통 거래집중도(시가)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 시가 거래집중도의 곱}$
10	공통 거래집중도(종가)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 종가 거래집중도의 곱}$

〈표 2〉는 호가자료를 이용한 10개 지표들에 대한 설명과 구체적인 계산방법을 설명한다. 첫 번째 ‘공통 종목_호가일(전체)’ 연계지표는 식 (1)의 ‘공통 종목_거래일’ 지표와 개념이 동일하며, 다만 거래일이 아니라 호가일을 사용하였다는 차이만 존재한다. 두 계좌가 동일 종목일 및 동일 거래일에 호가한 수가 얼마나 많은지에 대해 측정하는 연계지표라 할 수 있다. 두 번째 ‘공통 호가금액관여율’ 연계지표도 식 (3)의 ‘공통 거래금액관여율’과 동일한 개념이며,

거래금액이 아니라 호가금액을 사용하였다는 점만 차이가 난다. 마지막 세 개의 ‘공통 호가집중도 (전체/시가/종가)’ 연계지표도 식 (4)의 ‘공통 거래집중도’ 지표와 동일한 개념이며, 거래금액 대신 호가금액을 사용한 것만 다르다.

호가지표에서 새롭게 도입된 것은 3~5번째의 ‘공통 호가취소율(전체/매수/매도)’ 연계지표와 6~7번째의 ‘공통 호가공격성(매수/매도)’ 연계지표이다. 먼저 두 계좌 사이의 ‘공통 전체 호가취소율’ 연계지표는 다음 식 (5)와 같이 정의된다.

$$\begin{aligned} \text{공통 전체 호가취소율} &= \sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 전체호가취소율의 곱} \\ \text{계좌 } i \text{의 호가취소율} &= \frac{\text{계좌 } i \text{의 종목 당일호가취소금액}}{\text{계좌 } i \text{의 종목 당일호가금액}} \end{aligned} \quad (5)$$

특정 계좌에 대해 해당 종목의 당일 호가금액 중에 취소한 호가의 비중을 측정하며, 두 계좌에 대해 이 비중을 곱하였다. 이 곱을 공통거래일에 걸쳐서 합산하여서 연계지표를 구하였다.

〈표 2〉 호가자료를 이용한 계좌쌍 지표

번호	연계지표	설명
1	공통 종목_호가일(전체)	$\frac{\text{공통 종목 호가일}}{\text{Min}(i \text{계좌의 종목 호가일}, j \text{계좌의 종목 호가일})}$
2	공통 호가금액관여율(전체)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 호가관여율의 곱}$ $\text{호가관여율} = \frac{\text{계좌 } i \text{의 해당종목에 대한 호가거래금액}}{\text{해당종목의 총 호가거래금액}}$
3	공통 호가취소율(전체)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 전체호가취소율의 곱}$ $\text{계좌 } i \text{의 호가취소율} = \frac{\text{계좌 } i \text{의 종목 당일호가취소금액}}{\text{계좌 } i \text{의 종목 당일호가금액}}$
4	공통 호가취소율(매수)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 매수호가취소율의 곱}$
5	공통 호가취소율(매도)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 매도호가취소율의 곱}$
6	공통 호가공격성(매수)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 매수호가 공격성의 곱}$ $\text{매수호가공격성} = \frac{\text{매수호가가격} - \text{최우선매수호가}}{\text{최우선매수호가}} \times 100$
7	공통 호가공격성(매도)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 매도호가 공격성의 곱}$ $\text{매도호가공격성} = \frac{\text{매도호가가격} - \text{최우선매도호가}}{\text{최우선매도호가}} \times 100$
8	공통 호가집중도(전체)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 호가집중도의 곱}$ $\text{계좌 } i \text{의 호가집중도} = \frac{\text{계좌 } i \text{의 해당종목의 당일호가금액}}{\text{계좌 } i \text{의 당일호가금액}}$
9	공통 호가집중도(시가)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 시가 호가집중도의 곱}$
10	공통 호가집중도(종가)	$\sum_{t=\text{공통일}} \text{계좌 } i, j \text{의 종가 호가집중도의 곱}$

연계지표의 값이 클수록 해당 종목에 대해 두 계좌들이 호가를 취소하는 비율이 높음을 의미한다. 이는 많은 주가조작이 고가 매수주문 후 취소되는 방식의 허수주문에 의해 이루어지고 있음에서 착안하였다.

10개의 호가지표들 중에서 가장 특징적인 연계지표는 마지막 소개할 ‘공통 호가공격성’ 지표이다. 두 계좌 사이의 ‘공통 호가공격성’ 연계지표는 다음 식 (6)과 같이 정의된다.

$$\begin{aligned} \text{공통 매수 호가공격성} &= \sum_{i=\text{공통일}} \text{계좌 } i, j \text{의 매수 호가공격성의 곱} \\ \text{계좌 } i \text{의 매수 호가공격성} &= \frac{\text{계좌 } i \text{의 매수 호가가격} - \text{최우선매수호가}}{\text{최우선매수호가}} \times 100 \end{aligned} \quad (6)$$

특정 계좌의 매수 호가공격성은 계좌가 제출한 매수호가가격과 최우선매수호가와 차이를 최우선매수호가로 표준화한 값으로 측정하였다. 즉, 호가를 제출할 당시의 최우선매수호가에 비해 투자자가 상대적으로 높은 매수가격을 호가할수록 더 공격적인 주문으로 해석하였다. 이런 주문들이 많을수록 그 계좌는 큰 호가공격성을 갖게 된다. ‘공통 매수 호가공격성’ 연계지표는 두 계좌의 호가공격성을 곱함으로써 두 계좌의 호가공격성이 클수록 더 큰 값을 갖게 되며, 이 곱을 공통 거래일에 걸쳐서 합산하였다. 연계지표의 값이 클수록 해당 종목에 대해 두 계좌들이 공격적인 매수호가를 제출함을 의미한다. 매도호가에 대해서도 동일한 방법으로 연계지표를 정의하였다.

3.2 빅데이터 분석기법

빅데이터 분석을 위해서는 저장, 정제, 분석 및 가시화, 그리고 예측 기술 같은 분석 과정에 대한 이해가 필요하다. 첫째, 저장기술은 저장 단가의 절감, 자료 저장 및 검색 속도의 향상, 저장 신뢰도 및 안정성의 보장 등이 가장 큰 이슈이다. 증권거래 자료의 경우는 이미 증권 거래시스템을 통해서 자료가 수집되어 저장되고 있기 때문에 이와 관련하여 큰 문제는 없다고 생각된다.

둘째, 빅데이터 정제 기술은 분석 가능한 형태로 원시데이터를 정제하는 작업을 의미한다. 비정형 데이터(unstructured data, irregular formatted data)는 컴퓨터가 바로 처리할 수 없는 데이터이기 때문에 분석에 맞는 형태로 변환해야 한다. 정형 데이터(structured data)는 컴퓨터가 바로 처리할 수 있는 데이터이지만, 값의 누락이나 오류가 있기 때문에 점검이 필요하다. 증권거래 자료의 경우 정형 데이터로 컴퓨터가 바로 처리할 수 있기 때문에 큰 어려움은 없었다.

셋째, 빅데이터 분석 및 시각화 기술은 자료를 분석하여 그 분석결과를 구체적으로 보여주는 과정을 의미한다. 빅데이터 분석 기술은 결국 통계학에 속하는 부분이 된다. 통계학 중 데이터 마이닝 분야가 가장 많이 활용된다. 가트너 그룹은 데이터 마이닝을 통계, 수학, 그리고 패턴인식 등의 기술을 사용하여 대용량의 데이터 속에서 의미 있는 새로운 상관관계, 패턴, 트렌드 등을 발견하는 기술로 정의하고 있다.

넷째, 빅데이터 기반 예측 기술은 전체 데이터를 모두 조사해서 패턴별로 감시하다가 유사 패턴이 나오면 발생 사건을 예측한다.

본 연구의 대상이 되는 증권거래 빅데이터의 경우, 이미 정제되어 저장되기 때문에 첫 번째와 두 번째 과정은 큰 문제가 되지 않는다. 결국 세 번째에 해당하는 어떤 분석기법을 적용할지가 관건이다. 일반적으로 빅데이터의 분석 기술에는 텍스트 마이닝, 평판 분석, 클러스터 분석, 사회네트워크 분석 등의 4가지로 나누어 볼 수 있다.

첫째, 텍스트 마이닝(Text Mining)은 자연어 처리 기술을 이용하여 비·반정형 텍스트 데이터로부터 유용한 정보를 추출하거나 가공하는 기술이다. 이 기술을 통해 방대한 텍스트 속에 존재하는 의미 있는 정보를 추출해내고, 그 정보들의 연계성을 파악한다. 프로그래밍을 사용하여 인간이 사용하는 언어인 자연어를 분석하고 그 안에 숨겨진 의미를 발굴해 내는 것은 쉬운 작업이 아니다. 이를 위해 대용량 언어자원과 통계적, 규칙적 알고리즘이 사용되며, 주로 문서 분류, 문서 군집, 정보 추출, 문서요약 등에 응용된다.

둘째, 평판분석(Opinion Mining)은 웹사이트와 SNS 상에 나타난 여론과 의견을 분석하여 유용한 정보로 재가공하는 기술이다. 이 기술을 활용하면 네티즌들이 각각의 사건이나 뉴스에 대하여 이야기하는 댓글이나 포스팅 등을 긍정 또는 부정으로 분류하여 더욱 더 객관적이고 정확하게 평판을 파악할 수 있다. 아울러 특정 서비스 및 상품에 대한 시장규모 예측, 소비자의 반응, 입소문 분석 등에 도 활용할 수 있다.

셋째, 군집 분석(Cluster Analysis)은 각 개체의 유사성을 측정하여 서로 유사성이 높은 집단을 분류하고, 특정 집단에 속한 개체들의 유사성과 다른 집단에 속한 개체 간의 상이성을 규명하는 통계 분석 방법이다. 비슷한 특성을 가진 개체를 합쳐 가면서 최종적으로 유사 특성의 그룹을 발굴하는데 사용된다. 예를 들어 웹 사이트 상에서 주로 사진에 대해 이야기하는 사용자 그룹이 있을 수 있고, 음악에 대해 관심 있는 사용자 그룹이 있을 수 있다. 사진에 대해 관심있는 그룹은 카메라 등의 구매에 더 적극적일 것이며, 음악에 관심있는 그룹은 음향장비의 구매에 더 적극적일 것이다. 이러한 관심사나 취미에 따른 사용자 그룹을 군집분석을 통해 분류할 수 있으며 마케팅 등에 활용할 수 있다.

넷째, 사회네트워크 분석(Social Network Analysis)은 네트워크 연결구조 및 연결강도 등을 바탕으로 사용자의 명성 및 영향력을 측정한다. 사회네트워크 상에서 입소문의 중심이나 주된 연결 역할을 하는 사용자를 찾는데 주로 활용된다. 이렇게 사회네트워크 상에서 영향력이 있는 사용자를 인플루언서(influencer)라고 부르는데, 인플루언서 모니터링 및 관리는 마케팅 관점에서 대단히 중요하다.

4개의 분석기법들 중에서 본 연구에서 사회네트워크 분석을 사용하였다. 이에 대해 더 자세히 살펴보고 연계계좌분석에 적용할 수 있는 방안을 생각해보자. 사회네트워크 분석은 개인, 집단을 분석단위로 하지 않고, 개인과 개인의 연결 즉, 네트워크를 분석단위로 하는데 그 특징이 있다. 점(node)과 연결(connection)이 존재하는 네트워크이라면, 분석이 가능하며, 점 하나 또는 점의 집합에 대해서만 분석하던 기존 연구방법론에서 발견하지 못하는 현상을 분석할 수 있다는 것이 장점이다. 생물학, 범죄학, 서지학, 조직학, 인류학 등 다양한 분야에서 단백질 네트워크, 테러리스트 네트워크, 지식 네트워크, 조직 네트워크, 위계 네트워크 등을 연구하는데 활용되고 있다.

특히, 사회네트워크 분석은 불공정거래나 범죄행위를 적출하는데 효과적인 방법론이 될 수 있다. 불공정거래나 범죄행위와 같이 연결구조를 노출시키지 않으려는 네트워크에서는 매우 복잡한 차원의 연결구조를 가지는 것이 일반적이어서 사회네트워크 분석의 활용도가 더욱 높다. 예를 들어, 대포폰의 적출, 인터넷경매시장에서 허위로 경매가를 올리는 행위 적출, 테러리스트 적출 등의 분야에서 사회네트워크 분석이 활용되고 있다.

사회네트워크 분석에서 네트워크의 구조를 파악하는데 사용되는 지표는 다음과 같다. 첫째, 연결정도(degree)는 네트워크 전체의 연결(connection)의 수이다. 연결정도가 클수록 연계계좌군의 규모가 크면서 연계성이 큼을 의미한다. 둘째, 밀도(density)는 네트워크 전체의 연결수와 모든 가능한 연결수의 비율이다. 밀도가 클수록 연계계좌군의 연계성이 큼을 의미한다. 셋째, 다선중심성(degree centrality)은 한 점의 연결수와 한 점에서 모든 가능한 연결수의 비율이다. 선중심성이 클수록 해당 계좌가 연계계좌군에서 직접적인 연계성이 큼을 의미한다. 넷째, 매개중심성(between centrality)은 네트워크에서 매개자 혹은 중재자의 역할을 측정하는 지표이다. 매개중심성이 클수록 해당 계좌가 한 연계계좌군과 다른 연계계좌군의 매개자 역할을 하고 있음을 의미한다. 다섯째, 위계중심성(eigenvector centrality)은 다선중심성으로부터 발생하는 연계성과 자신과 연결된 다른 점(node)의 연계성을 합하여 측정하는 지표이다. 자신의 다선중심성과 매개중심성을 낮더라도 다선중심성과 매개중심성이 높은 점(node)와 연결되어 있는 경우, 위계중심성은 높다. 위계중심성이 높을수록 네트워크 상에서의 연계성이 높음을 의미한다.

본 연구는 주식시장에서 연계계좌군을 적출하는 것을 목표로 하고 있기 때문에 혐의계좌들 사이에서도 어떤 하부구조(sub-structure)가 존재하는지도 관심을 가져야 한다. 하부구조를 파악하는 데 있어서 여러 방법들이 있지만, 본 연구에서는 K-코어(K-core) 방법론을 사용하였다. K-코어 알고리즘은 가중치가 낮은 연결과 노드를 단계적으로 제거하여, 복잡한 네트워크의 의미를 분석하는데 핵심되는 하부구조를 얻는 것이다. 네트워크 상의 노드들이 k 이상의 연결수를 갖는 경우, 코어 값을 k 로 정의하게 된다. 즉, $(k+1)$ -코어 네트워크는 $(k+1)$ 의 연결수를 가지는 노드들로 구성된 네트워크로 이 네트워크가 형성될 때까지 연결수가 k 이하인 노드를 제거하는 과정을 반복한다. 가중치가 낮은 연결과 노드를 단계적으로 제거함으로써 핵심부분을 파악하는데 도움을 준다. 이와 같이 K-코어 분석을 적용하면, 가장 핵심이 되는 노드를 중심으로 네트워크가 결집되어 몇 개의 핵심그룹이 나타나게 된다.

4. 자료

본 연구를 위해서 한국거래소에서 제공하는 주식에 대한 거래자료 및 호가자료를 사용하였다. 주식시장은 유가증권, 코스닥, 코넥스 시장이 해당되며, 기간은 2017년 3월부터 5월까지 3개월이다. 거래자료에는 해당 주식의 모든 거래에 대한 거래시간, 거래가격, 거래량 및 매도자와 매수자에 대한 정보들이 포함되어 있다. 거래자에 대한 정보는 개인, 기관, 외국인 등의 분류 및 국적, 임의번호로 처리된 계좌번호 등이다. 호가자료에는 해당 주식의 모든 호가의 호가시간, 주문종류, 주문가격 및 주문자에 대한 정보가 포함되어 있다.

<표 3> 패널 A는 거래자료를 통해서 3개월 동안 거래에 참여한 모든 계좌들의 요약통계량을 보여주고 있다. 표본기간 동안 총 2,973,780개의 계좌가 거래에 참여하였다. 계좌들의 거래량 평균이 49,912주이며, 거래금액 평균은 약 3억 7천만 원이다. 이 중 프로그램 거래비중의 평균은 0.5%, 시가거래 비중은 1.88%, 종가거래 비중은 2.96%에 이른다. 시가거래보다는 종가거래가 더 활발함을 알 수 있다.

패널 B는 호가자료를 통해서 3개월 동안 호가를 제출한 적이 있는 모든 계좌들의 요약통계량을 제시하고 있다. 분석 결과 총 3,080,051개의 계좌가 표본기간 동안 호가를 제출하여, 거래계좌보다 약 10만 개가 더 많다. 주의할 점은 시장가 주문의 경우 가격이 0으로 표기되어 있어 호가금액이 0인 계좌가 요약통계량에 존재한다는 것이다. 계좌들의 호가량 평균이 952,302주이다. 이중

<표 3> 전체 계좌의 요약통계량

패널 A: 전체 거래 계좌

	계좌수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
거래량(주)	2,973,780	49,912	1,411,413	1	224	1,880	13,088	1,218,411,522
거래금액(천원)	2,973,780	374,048	18,087,724	0	3,715	19,905	100,430	13,972,761,374
프로그램 거래량(주)	2,973,780	1,649	455,286	0	0	0	0	457,054,032
프로그램 거래비중(%)	2,973,780	0.50	6.78	0.00	0.00	0.00	0.00	100.00
시가 거래량(주)	2,973,780	494	8,403	0	0	0	0	4,561,755
시가 거래비중(%)	2,973,780	1.88	10.47	0.00	0.00	0.00	0.00	100.00
종가 거래량(주)	2,973,780	1,046	30,543	0	0	0	8	24,872,231
종가 거래비중(%)	2,973,780	2.96	11.90	0.00	0.00	0.00	0.15	100.00

패널 B: 전체 호가 계좌

	계좌수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
호가량(주)	3,080,051	952,302	235,450,674	1	357	3,300	27,405	300,392,667,884
호가 취소량(주)	3,080,051	162,898	39,924,756	0	0	1	1,748	29,962,385,578
호가 호가율(%)	3,080,051	7.16	10.62	0.00	0.00	0.06	12.02	50.00
시가 호가량(주)	3,080,051	16,313	1,270,580	0	0	0	246	1,454,686,100
시가 호가비중(%)	3,080,051	7.02	18.54	0.00	0.00	0.00	2.66	100.00
종가 호가량(주)	3,080,051	15,796	3,225,668	0	0	0	45	3,005,921,828
종가 호가비중(%)	3,080,051	2.68	10.55	0.00	0.00	0.00	0.44	100.00

호가취소량 평균은 162,898주로써, 호가취소율의 평균은 7.16%에 이른다. 평균 시가호가량은 16,312주로 전체 호가량 대비 비중은 7.02%이며, 종가호가량은 15,796주로, 평균 비중은 2.68%이다. 시가호가량이 종가호가량보다 더 많은데 이는 거래량과는 반대되는 현상이다.

연계계좌에 대한 본격적 연구는 전체 계좌를 사용하기 보다는 표본을 추출하여 분석을 진행하였다. 거래에 참여한 계좌가 총 2,954,495개인 상황에서, 전체 계좌를 사용하여 2개씩의 계좌쌍을 분석한다고 할 때 계산량이 너무 많은 어려움이 있다. 따라서 실제 분석에서는 전체 거래계좌들

〈표 4〉 표본계좌의 요약통계량

패널 A: 표본계좌의 거래 요약통계량

	계좌수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
거래량(주)	2,000	43,607	237,284	1	230	1,776	12,485	6,130,724
시가	2,000	353	3,231	0	0	0	0	100,000
거래량(주)								
종가	2,000	720	5,871	0	0	0	6	197,674
거래량(주)								
시가	2,000	1.89	10.27	0.00	0.00	0.00	0.00	100.00
거래비중(%)								
종가	2,000	3.13	13.05	0.00	0.00	0.00	0.08	100.00
거래비중(%)								
종목_거래일수	2,000	29	77	1	2	6	24	1,608
종목_시가	2,000	1	3	0	0	0	0	64
종목_종가	2,000	1	5	0	0	0	1	96
종목_거래일수								

패널 B: 표본계좌의 호가 요약통계량

	계좌수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
호가량(주)	2,000	114,955	685,132	1	406	3,324	28,008	15,972,254
취소								
호가량(주)	2,000	22,071	189,226	0	0	3	1,684	6,286,082
시가								
호가량(주)	2,000	9,093	58,409	0	0	0	342	1,238,984
종가								
호가량(주)	2,000	3,435	47,107	0	0	0	50	1,863,405
취소								
호가비중(%)	2,000	7.00	10.21	0.00	0.00	0.21	11.88	48.65
시가								
호가비중(%)	2,000	7.01	17.86	0.00	0.00	0.00	3.26	100.00
종가								
호가비중(%)	2,000	2.78	11.15	0.00	0.00	0.00	0.49	100.00
종목_호가								
종목_거래일수	2,000	39	94	1	3	9	34	1,609

중에서 다음과 같은 순서로 임의추출한 2,000개의 계좌를 분석하는 방법을 택하였다. 첫째, 3개월 동안 거래에 참여한 전체 계좌들 중에서 프로그램매매에 참여한 계좌들을 제외하였다. 프로그램매매는 컴퓨터 시스템을 통하여 자동으로 거래가 이루어지기 때문에 의도적인 시세조종과는 관계가 없다고 생각하여 분석대상에서 제외하였다. 이러한 필터링을 적용하였을 때, 총 2,954,495개의 계좌가 남았다. 둘째, 이들 계좌들 중에서 종목_거래일_횟수 순으로 정렬하여서 2,000개를 추출함으로써 다양한 유동성의 주식들이 포함될 수 있도록 하였다.

<표 4>는 최종 선택된 2,000개의 표본계좌에 대한 요약통계량을 제시해주고 있다. 계좌들의 거래량 분포가 전체 계좌의 거래량 분포를 보여주는 <표 3>과 크게 다르지 않다. 패널 A의 거래에 대한 요약통계량을 보면, 계좌들의 거래량 평균이 43,607주이며, 시가 거래비중은 평균 1.89%, 종가 거래비중은 평균 3.13%이다. 종목_거래일 수는 평균 29거래일이다. 패널 B는 호가에 대한 요약통계량을 보여준다. 계좌들의 평균 호가량은 114,955주로 거래량보다 3배 정도 크다. 호가량의 평균이 <표 3> 패널 B의 전체표본의 평균 952,302보다 작지만, 중간값 3,324는 전체표본의 중위값 3,300과 비슷하다. 전체표본에 극단적 이상치들이 존재하지만 오히려 표본계좌에는 이런 계좌들이 제거되면서도 전체 분포는 비슷함을 알 수 있다. 취소 호가비중은 7.01%, 시가 호가비중은 7.01%, 종가 호가비중은 2.78%로 전체표본과 비슷하다. 이런 점들은 표본계좌가 전체표본을 대표하기에 부족하지 않음을 보여준다.

5. 분석 결과

5.1 1단계: 계좌쌍의 연계지표

먼저 각 계좌쌍의 연계지표를 <표 1>과 <표 2>에 정의에 따라서 계산하였다. 각 거래지표에 따른 계좌쌍들의 지표의 분포를 <표 5>에 보고하였다. 표본이 총 2,000개의 계좌이기 때문에 계좌쌍은 $2,000 \times 1,999 / 2 = 1,999,000$ 계좌쌍이 존재한다. 하지만 연계지표 계산을 통해서 값이 0인 경우는 제거하고 실제 관계가 나타나는 계좌쌍들에 대한 분포를 분석하였다. 10개 지표를 보고하는 것은 분량이 너무 많아 이후 분석에서 주로 사용될 3개 지표만 보고하였다.

첫 번째 거래 연계지표인 ‘공통 종목_거래일(전체) 지표’는 두 계좌가 같은 종목과 같은 거래일을 가지고 있는 횟수를 계산한 것으로 상당히 광범위하게 두 계좌 간의 관계를 나타내 준다. <표 5>의 패널 A를 보면, 최상위 10%인 3,978개의 계좌쌍은 지표값이 0.527이며, 3사분위 이상의 계좌쌍은 1의 값을 갖는다.

패널 B의 ‘통정 종목_거래일 지표’는 두 계좌가 해당 종목에서 비슷한 수량의 순매도와 순매수 거래량을 보이며 해당일을 통정거래일로 인식한 후 이 통정 종목_거래일의 횟수를 계산하였다. 이 지표는 두 계좌 간의 관계를 엄격하게 보여주는 지표이다. 즉, 두 계좌간의 관계가 나오기는 쉽지 않지만, 나온다면 아주 밀접한 관계가 있다고 해석할 수 있다. 분포를 보면, 최상위 1,587개의 계좌쌍은 평균 통정 종목_거래일이 7일이며, 이 중 최댓값은 149종목_거래일에 이른다.

패널 C ‘공통 거래집중도(전체) 지표’는 계좌가 해당 종목의 해당일 거래량을 계좌의 해당일 총 거래량으로 나눈 값을 집중도로 정의된다. 이는 계좌가 해당 종목을 어느 정도 집중해서

거래하였는가를 보여준다. 이 지표는 시세조종 계좌일수록 소수의 종목에 집중하는 데서 착안된 것이다. 최종적으로 해당 지표는 두 계좌의 거래집중도를 곱하고 이를 종목_거래일에 대해 합산하여 계산한다. 따라서 지표값이 클수록 두 계좌 간의 종목에 대한 집중도가 강함을 의미한다. 전체 거래집중도의 경우, 최상위 4,889개의 계좌쌍은 1.0 이상의 값을 갖는다. 거래집중도가 1이 넘어가는 것은 공통 거래일수에 따라 합하기 때문이다. 중간값이 1.085, 평균값이 1.399로 분포가 거의 대칭적이다.

〈표 5〉 거래 연계지표의 분포

패널 A: 공통 종목_거래일(전체) 지표의 분포

	계좌쌍수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
1(최하위 10%)	3,794	0.011	0.003	0.003	0.009	0.011	0.013	0.015
2	3,793	0.019	0.002	0.015	0.017	0.019	0.021	0.023
3	3,942	0.027	0.002	0.023	0.025	0.027	0.029	0.031
4	3,495	0.036	0.003	0.031	0.034	0.036	0.038	0.041
5	4,082	0.047	0.004	0.042	0.043	0.047	0.050	0.053
6	3,662	0.061	0.005	0.053	0.057	0.063	0.066	0.071
7	3,818	0.081	0.007	0.071	0.075	0.080	0.087	0.091
8	3,744	0.113	0.012	0.091	0.100	0.111	0.125	0.142
9	3,635	0.175	0.026	0.143	0.154	0.167	0.200	0.246
10(최상위 10%)	3,978	0.527	0.300	0.250	0.267	0.373	1.000	1.000

패널 B: 통정 종목_거래일 지표의 분포

	계좌쌍수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
1(최하위)	7,467	1	0	1	1	1	1	1
2	1,509	2	0	2	2	2	2	2
3(최상위)	1,587	7	10	3	3	4	7	149

패널 C: 공통 거래집중도(전체) 지표의 분포

	계좌쌍수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
1(최하위 10%)	4,070	0.002	0.002	0.000	0.000	0.002	0.003	0.006
2	4,070	0.014	0.005	0.006	0.009	0.013	0.018	0.023
3	4,071	0.036	0.009	0.023	0.028	0.036	0.044	0.053
4	4,070	0.076	0.014	0.053	0.063	0.075	0.087	0.101
5	4,070	0.134	0.020	0.101	0.117	0.133	0.152	0.170
6	4,073	0.222	0.032	0.170	0.194	0.220	0.250	0.280
7	4,068	0.353	0.046	0.280	0.314	0.349	0.393	0.440
8	4,071	0.550	0.070	0.440	0.491	0.541	0.606	0.682
9	3,251	0.832	0.093	0.682	0.750	0.826	0.911	1.000
10(최상위 10%)	4,889	1.399	0.724	1.000	1.000	1.085	1.526	12.084

다음은 표본 2,000계좌의 계좌쌍들에 대한 호가 연계지표를 살펴보고자 한다. 각 호가지표에 따라 계좌쌍들의 연계지표의 분포를 살펴보았으며, 계좌쌍의 값이 0인 경우는 제거하였다. 여기서도 10개 모두 보다는 이후에 많이 활용되는 5개 지표만 보고하였다.

패널 A의 ‘공통 종목_호가일(전체) 지표’는 두 계좌가 같은 종목과 같은 호가일을 가지고 있는 횟수를 계산한 것으로 상당히 광범위하게 두 계좌 간의 관계를 보여준다. 최종 지표는 두 계좌 중 작은 종목_거래일 횟수로 나누어서 표준화하였다. 최상위 10%인 2,727개의 계좌쌍은 지표값 평균이 0.451이다. 3사분위 이상의 계좌쌍은 0.667 이상의 값을 가지며, 최댓값은 1이다.

패널 B의 ‘공통 호가취소율(전체) 지표’는 두 계좌가 각각 해당 종목의 해당일 호가량 중에서 취소한 비중을 계산한 후에 이 둘을 곱하였다. 해당 지표는 이 둘의 곱을 종목_거래일에 대해서 합산하였다. 지표값이 클수록 두 계좌 간의 관계가 강함을 의미한다. 분포를 보면, 공통 종목_호가 취소율(전체)에서 최상위 1,501개의 계좌쌍은 0.029 이상의 값을 가지며, 중간값 0.055, 최댓값 0.482로 값이 크지 않다.

패널 C/D의 ‘공통 호가공격성(매수/매도) 지표’는 매수(매도)호가 공격성은 지정가 매수(매도) 가격에서 최우선매수(매도)호가와의 차이를 최우선매수(매도)호가로 나눈 값으로 측정하였다. 매수의 경우 시세조종 혐의자는 가격을 올리기 위해서 높은 가격으로 호가를 내리는 경향을 보일 수 있다. 두 계좌의 공통 호가공격성은 각각의 호가공격성을 곱하여 구함으로, 두 계좌가 일치된 방향으로 공격적으로 호가를 제시한다면 호가공격성의 값이 커지며, 이는 연계계좌일 가능성이 높다고 판단하였다. 해당 지표는 최종적으로 이 둘의 곱을 종목_거래일에 대해서 합산하였다.

패널 C의 공통 매수호가 공격성 경우, 최상위 계좌쌍 그룹에서 중간값은 45.364이지만, 평균값은 3792.341이다. 이는 극단적으로 같은 방향으로 호가를 제출하는 계좌쌍들이 존재함을 의미한다.

〈표 6〉 호가 연계지표의 분포

패널 A: 공통 종목_호가일(전체) 지표의 분포

	계좌쌍수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
1(최하위 10%)	2,691	0.007	0.002	0.002	0.006	0.007	0.009	0.010
2	2,707	0.012	0.001	0.010	0.011	0.012	0.013	0.014
3	2,473	0.017	0.001	0.014	0.016	0.017	0.018	0.019
4	2,780	0.022	0.002	0.019	0.020	0.022	0.023	0.025
5	2,677	0.029	0.002	0.026	0.028	0.029	0.031	0.033
6	2,669	0.038	0.003	0.033	0.036	0.037	0.040	0.042
7	2,742	0.051	0.005	0.042	0.046	0.050	0.056	0.059
8	2,587	0.072	0.008	0.059	0.065	0.071	0.077	0.090
9	2,603	0.116	0.020	0.090	0.100	0.111	0.133	0.163
10(최상위 10%)	2,727	0.451	0.320	0.167	0.200	0.286	0.667	1.000

패널 B: 공통 호가취소율(전체) 지표의 분포

	계좌쌍수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
1(최하위)	1,501	0.000	0.000	0.000	0.000	0.000	0.001	0.001
2	1,501	0.002	0.001	0.001	0.001	0.002	0.003	0.004
3	1,502	0.006	0.002	0.004	0.005	0.006	0.008	0.010
4	1,501	0.017	0.005	0.010	0.013	0.017	0.022	0.029
5(최상위)	1,501	0.072	0.049	0.029	0.038	0.055	0.087	0.482

패널 C: 공통 호가공격성(매수) 지표의 분포

	계좌쌍수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
1(최하위 10%)	1,852	-61.774	732.431	-20,378	-9.035	-4.568	-2.747	-1.805
2	1,853	-0.986	0.375	-1.804	-1.282	-0.907	-0.659	-0.484
3	1,853	-0.272	0.099	-0.484	-0.345	-0.260	-0.183	-0.133
4	1,852	-0.067	0.032	-0.133	-0.094	-0.061	-0.038	-0.022
5	1,853	0.003	0.013	-0.021	-0.008	0.002	0.013	0.026
6	1,853	0.079	0.040	0.026	0.044	0.070	0.110	0.163
7	1,852	0.384	0.165	0.163	0.237	0.348	0.509	0.741
8	1,853	1.603	0.649	0.741	1.041	1.457	2.107	2.997
9	1,853	7.138	3.226	3.001	4.324	6.301	9.523	14.564
10(최상위 10%)	1,852	3792.341	102,198	14.571	24.013	45.364	98.057	3,847,109

패널 D: 공통 호가공격성(매도) 지표의 분포

	계좌쌍수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
1(최하위 10%)	2,195	-4.722	24.645	-905.349	-3.258	-1.641	-0.990	-0.661
2	2,195	-0.370	0.130	-0.661	-0.466	-0.340	-0.260	-0.193
3	2,196	-0.120	0.036	-0.193	-0.149	-0.114	-0.088	-0.067
4	2,195	-0.039	0.014	-0.067	-0.050	-0.037	-0.026	-0.016
5	2,196	-0.002	0.008	-0.016	-0.009	-0.001	0.003	0.011
6	2,195	0.030	0.013	0.011	0.019	0.028	0.041	0.056
7	2,196	0.106	0.035	0.056	0.075	0.101	0.134	0.177
8	2,195	0.320	0.098	0.178	0.234	0.310	0.398	0.521
9	2,196	1.109	0.452	0.521	0.722	0.999	1.431	2.170
10(최상위 10%)	2,195	18.091	45.525	2.171	3.497	6.540	15.770	976.184

패널 E: 공통 호가집중도(전체) 지표의 분포

	계좌쌍수	평균	표준편차	최솟값	1사분위	중간값	3사분위	최댓값
1(최하위 10%)	5,353	0.001	0.001	0.000	0.000	0.001	0.002	0.004
2	5,353	0.009	0.003	0.004	0.006	0.009	0.012	0.016
3	5,354	0.026	0.006	0.016	0.020	0.025	0.031	0.038
4	5,353	0.055	0.011	0.038	0.046	0.054	0.065	0.076
5	5,355	0.104	0.018	0.076	0.089	0.103	0.119	0.138
6	5,352	0.183	0.029	0.138	0.159	0.180	0.206	0.238
7	5,353	0.308	0.044	0.238	0.269	0.304	0.342	0.392
8	5,354	0.510	0.073	0.392	0.446	0.500	0.572	0.651
9	4,432	0.812	0.102	0.651	0.723	0.805	0.900	1.000
10(최상위 10%)	6,274	1.638	1.504	1.000	1.000	1.209	1.818	45.000

패널 D는 공통 매도호가 공격성 경우, 최상위 계좌쌍 그룹에서 중간값은 6.540이지만, 평균값은 18.091이다. 같은 방향으로 호가를 제출하는 계좌쌍들이 존재함을 보여주지만, 그 강도에 있어서는 매수호가 공격성보다 약함을 보여준다.

패널 E의 ‘공통 호가집중도(전체) 지표’는 계좌가 해당 종목의 해당일 호가량을 계좌의 해당일 총 호가량으로 나눈 값을 집중도로 정의한다. 이는 계좌가 해당 종목을 어느 정도 집중해서

호가를 제출하였는가를 보여준다. 이는 시세조종 계좌일수록 소수의 종목에 집중하는 데서 착안한 것이다. 최종 연계지표는 두 계좌의 호가집중도를 곱하고 이를 종목_거래일에 대해 합산하였다. 따라서 지표값이 클수록 두 계좌 간의 종목에 대한 집중도가 강함을 의미한다. 분포를 보면, 최상위 6,274개의 계좌쌍은 1.0 이상의 값을 갖는다. 중간값이 1.209, 평균값이 1.638로 분포가 거의 대칭적이다.

5.2 2단계: 연계지표의 결합 또는 사건일을 통한 특정

앞 장에서는 각 계좌쌍에 대해 20개의 연계지표를 계산하여 계좌 간의 연관성을 측정하는 방법을 소개하였다. 이 부분에서는 전체 계좌쌍을 필터링하여 혐의 연계계좌쌍을 찾는 방법을 제안하고자 한다. 이를 위해서 앞에서 구한 연계지표들을 결합하는 방식과 사건일을 통해 특정하는 방법 2가지를 고려하였다.

5.2.1 연계지표의 결합

여기서는 2개 이상의 연계지표를 결합하여 혐의 연계계좌쌍을 특정에 사용하는 방법을 살펴보고자 한다. 어떤 연계지표를 결합할 것인지는 분석하고자 하는 불공정거래의 형태에 따라서 다수의 지표를 선택하여 결합할 수 있다. 예를 들어, 주가를 끌어 올리려는 주가조작자는 공격적인 매수주문이나 허수 매수주문을 사용할 가능성이 많다. 이는 연계지표들 중에서 호가 공격성이나 호가취소를 지표들과 연관되어 있으며 이 연계지표 2개를 결합하여 사용할 수 있다. 또한 주가상승 후 이익실현을 위해 대량으로 매도해야 할 필요가 있는데, 이를 위해 시가나 종가시점을 이용할 수 있는 것이 용이할 수 있다. 이를 추적하기 위해서는 시가나 종가 거래지표 2개를 결합하여 살펴볼 수 있을 것이다. 여기서는 위의 방법을 문헌에서 가장 많이 언급되는 주가조작의 형태인 Pump and Dump, 통정매매, 시가 및 종가시세 조종 등의 불공정거래에 적용해 보고자 한다.

첫째, 대표적인 불공정 거래 행위인 ‘Pump and Dump’에 대해 연계지표 결합 방법을 사용하였다. Pump and Dump는 매입한 주식의 가격을 폭등시킨 뒤 팔아치우는 불공정거래 행위이다. Pump는 주가를 끌어올리기 행위로 이를 위해서 공격적인 매수주문이나 허수 매수주문을 낼 수 있다. 이런 행위를 하는 계좌를 찾아내기 위해서는 공통 호가공격성이나 호가취소를 지표로 접근할 수 있다. 다른 사람보다 높은 가격의 매수 주문을 계속적으로 내서 주가를 의도적으로 상승시키는 경우는 앞 장에서 제안한 호가공격성을 통해서 측정가능하다. 허수성 주문은 실제로 매수나 매도할 의사가 없이 다른 사람의 매수나 매도를 유인하기 위해 체결가능성이 낮은 주문을 대량으로 제출하거나, 주문을 낸 후 아주 짧은 시간에 정정 또는 취소를 반복하는 경우이다. 이런 경우는 본 연구에서 제안한 공통 호가취소를 지표로 추적할 수 있을 것이다. Dump는 주가상승 후 주식을 매도하는 행위를 지칭하며, 일반적으로 대량 매도하기 위해서 시가나 종가시점을 이용할 수 있다.

<표 7>은 다양한 지표의 결합을 통하여 Pump and Dump 행위를 하는 연계계좌들을 추출하고자 시도하였다. 첫 행은 (공통 호가공격성 $\geq$ 80th percentile) 필터를 적용하여 호가 공격성이 큰 계좌들, 즉 Pump 가능성이 많은 계좌들을 찾아내고, 동시에 (공통 종목_거래일

>= 80th percentile) 필터를 적용하여 같은 거래일에 많이 거래하는 계좌들, 즉 Dump 가능성이 많은 계좌들을 찾아낸다. 이 두 지표에 의해서 걸러내진 계좌쌍은 총 228개이다. 이 계좌들은 Pump and Dump 행위 가능성이 높은 연계계좌군이라 할 수 있다. 둘째 행은 (공통 호가취소율 >= 60th percentile) 필터를 적용하여 호가취소율이 큰 계좌들로 Pump 가능성이 많은 계좌들을 찾아내고, 동시에 (공통 종목_거래일 >= 80th percentile) 필터를 적용하여 같은 거래일에 많이 거래하는 Dump 가능성이 많은 계좌들을 찾아내었다. 이 두 지표에 의해서 추출된 계좌쌍은 총 49개이다. 세 번째, 네 번째 행은 (거래 집중도 >= 80th percentile) 필터를 통하여 한 종목에 대한 거래집중도가 높은 계좌들을 Pump and Dump 행위 가능성이 높은 연계계좌군으로 인지하였다.

마지막 다섯 번째 행은 지금까지의 지표들을 모두 결합하였다((공통 호가공격성 >= 80th percentile) and (공통 종목_거래일 >= 80th percentile) and (거래 집중도 >= 80th percentile)). 계좌 간 호가 공격성이 높고, 공통 종목과 거래일에 많으며, 한 종목에 대한 거래 집중도가 높을수록 Pump and Dump 행위 가능성이 높은 연계계좌군이라고 인지하였다. 필터링 결과 총 105개의 계좌쌍을 찾아내었다.

연구자가 실제 이 방법을 적용할 때, 몇 가지 고려할 사항들이 있다. 첫째, 본 연구에서 제안한 연계지표 이외에 자신의 목표에 더 적합하다고 생각되는 다른 지표를 사용할 수도 있다. 논문에 소개된 연계지표로 충분하지, 아니면 자신의 연구목적에 더 적합한 측정 지표를 만들지는 연구자가 고민해야 할 부분이다. 둘째, 본 연구에서는 60th 또는 80th percentile 등의 절단점(cutoff point)을 사용하였지만 분석의 목적에 맞게 변경하여 적용할 수도 있다. 최적 절단점에 대한 연구는 향후 더 진행되어야 할 부분이다.

〈표 7〉 Pump and Dump 혐의 연계계좌 추적을 위한 연계지표의 결합

필터	계좌쌍수	공통 호가공격성 최솟값	공통 종목_거래일 최솟값	공통 매수 호가취소율 최솟값	거래 집중도 최솟값
(공통 호가공격성 >= 80th percentile) and (공통 종목_거래일 >= 80th percentile)	228	3.002	0.25		
(공통 호가취소율 >= 60th percentile) and (공통 종목_거래일 >= 80th percentile)	49		0.25	0.107	
(공통 호가공격성 >= 80th percentile) and (거래 집중도 >= 80th percentile)	838	3.012			0.682
(공통 호가취소율 >= 60th percentile) and (거래 집중도 >= 80th percentile)	267			0.105	0.684
(공통 호가공격성 >= 80th percentile) and (공통 종목_거래일 >= 80th percentile) and (거래 집중도 >= 80th percentile)	105	3.119	0.25	0	0.709

다음은 또 다른 불공정거래인 ‘통정거래’ 행위를 다루고자 한다. 통정거래는 거래가 성황리에 이루어지고 있는 것처럼 보이게 하여 일반인이 그릇된 투자판단을 하도록 유도하는 위장거래의

한 행태로, 사전에 약속한 특정 상대방과 주식을 매매 하는 것이다. 이를 찾아내기 위해 고안된 ‘통정 종목_거래일’ 지표를 기본적으로 사용할 수 있다. 통정 종목_거래일은 두 계좌가 특정 종목에 대해서 특정 거래일에 거의 동일한 금액을 한 계좌는 매수하고 한 계좌는 매도한 종목_거래일을 계수한다. 계좌쌍의 통정 종목_거래일이 3회 이상이면 통정계좌 혐의군의 가능성이 높은 것으로 인지하였다. 여기에 몇 가지 다른 연계지표를 사용한 필터링을 추가하여 연계계좌군의 범위를 좁히는 방향으로 지표를 결합하였다.

<표 8>의 첫 행은 (공통 종목_호가일 >= 80th percentile) 필터를 추가하여, 동일한 종목을 같은 날 호가할 가능성이 높을수록 더 불공정거래 혐의가 높다고 인식하였다. 결과적으로 총 421계좌쌍이 통정거래 계좌로 인지되었다. 둘째 행은 (공통 종목_거래일 >= 80th percentile) 필터를 추가하여, 동일한 종목을 같은 날 거래할 가능성이 높을수록 더 불공정거래 혐의가 높다고 인식하였다. 결과적으로 총 338계좌쌍이 추출되었다. 셋째 행은 공통 거래집중도가 1 이상인 계좌들을 혐의계좌로 인식하였으며, 총 496계좌쌍이 인지되었다. 마지막 넷째 행은 위의 지표들을 다 결합하여 통정 혐의계좌들을 추출하였다. 결과적으로 총 187계좌쌍이 인지되었다.

세 번째로 시가 및 종가 시세조종행위를 다루고자 한다. 이는 시가나 종가 형성시 제출된 주문가격이나 주문량이 하루 중의 주가수준을 결정하는데 기준점 역할을 하는 점을 이용하여 시세를 조작하려는 행위이다. 이를 위해 시가 결정시 전일 종가대비 고가 매수주문을 내거나, 종가 결정시 직전가 대비 고가 매수주문을 제출할 수 있다. 본 논문에서는 시가와 종가 시세조종 혐의계좌를 각각 나누어서 분석하였다. 처음에는 시가와 종가 두 시간대 모두 활동하는 혐의 연계계좌쌍을 찾으려고 하였으나 인지되는 계좌쌍들이 거의 없었다. 따라서, 시가와 종가를 나누어서 각각 분석하였다.

<표 8> 통정거래 혐의 연계계좌 추적을 위한 연계지표의 결합

필터	계좌쌍수	통정 종목_거래일	공통 종목_호가일 최솟값	공통 종목_거래일 최솟값	거래 집중도 최솟값
(통정 종목_거래일>= 3) and (공통 종목_호가일>= 80th percentile)	421	3	0.091		
(통정 종목_거래일>= 3) and (공통 종목_거래일>= 80th percentile)	338	3		0.144	
(통정 종목_거래일>= 3) and (공통 거래집중도>= 1)	486	3			1.000
(통정 종목_거래일>= 3) and (공통 종목_호가일>= 80th percentile) (공통 종목_거래일>= 80th percentile) (공통 거래집중도>= 1)	187	3	0.091	0.144	1.011

<표 9>에서 첫 번째, 두 번째, 네 번째 행은 시가 시세조종에 관한 것이며, 나머지 행은 종가 시세조종에 관한 것이다. <표 9>에서 일단 시가(종가) 시간대의 공통 종목_거래일 기준으로

필터링하였다. 즉, (공통 시가 종목_거래일 ≥ 0.25) 또는 (공통 시가 종목_거래일 ≥ 0.1) 기준을 적용하였다. 이를 통해 일단 시가(종가) 시간대에 동일한 종목, 같은 거래일을 가진 계좌쌍들을 찾아내었다. 이후 동시에 호가공격성이나 거래집중도 등의 지표도 같이 결합하였다. 예를 들어, 첫째 행의 시가 시세조종 혐의 계좌군은 (공통 시가 종목_거래일 ≥ 0.25) and (공통 호가공격성 ≥ 80 th percentile) 두 가지 필터링을 적용하였는데, 결과적으로 8개의 계좌쌍이 인지되었다.

〈표 9〉 시가 및 종가 시세조종 혐의 연계계좌 추적을 위한 연계지표의 결합

필터	계좌쌍수	공통 시가 종목_거래일 최솟값	공통 호가공격성 최솟값	공통 거래집중도 최솟값	공통 종가 종목_거래일 최솟값
(공통 시가 종목_거래일 ≥ 0.25) and (공통 호가공격성 ≥ 80 th percentile)	8	0.25	3.864		
(공통 시가 종목_거래일 ≥ 0.25) and (공통 거래집중도 ≥ 1)	31	0.25		1	
(공통 종가 종목_거래일 ≥ 0.25) and (공통 호가공격성 ≥ 80 th percentile)	12		3.126		0.333
(공통 종가 종목_거래일 ≥ 0.25) and (공통 거래집중도 ≥ 1)	22			1	0.25
(공통 시가 종목_거래일 ≥ 0.1) and (공통 거래집중도 ≥ 0.5)	53	0.1		0.522	
(공통 종가 종목_거래일 ≥ 0.1) and (공통 거래집중도 ≥ 0.5)	52			0.503	0.1

5.2.2 특정 사건일을 사용한 방법

지금까지의 분석은 전체 표본기간을 대상으로 연계지표를 계산하여 사용하였다. 하지만 모든 거래일이 다 의미있는 것은 아니다. 시세조종이나 미공개정보이용 등의 불공정거래가 발생할 가능성이 높은 거래일들을 중심으로 지표를 계산하면 혐의 연계계좌를 더 정확하게 찾아낼 수 있을 것이다.

다음과 같은 거래일들을 생각해 볼 수 있다. 수급에 의한 시세조종 가능성이 높은 거래일들은 상한가일, 가격 급등(락)일, 연속 거래일 동안 상승 또는 하락하는 경우 등을 생각할 수 있다. 상한가일의 예를 들어보자. 상한가일에는 가격을 상한가에 고정시키기 위해 주가조작자들이 과도하게 상한가 매수주문을 내는 ‘상한가 고정’ 행위가 존재할 수 있다. 분석을 위해, 먼저 표본기간 동안의 상한가일을 확인하였다. 표본기간 2017년 3월부터 5월까지 동안 일별 수익률이 14%를 초과하는 종목_거래일을 추출하였을 때, 총 20,523개의 종목_거래일이 존재하였다. 위의 종목_거래일 중에 상한가 고정행위를 하는 계좌쌍들을 추출하기 위해 여러 연계지표를 사용한 필터링을 시행하였다.

〈표 10〉은 그 결과를 보여주고 있다. 〈표 10〉의 첫 번째 행은 상한가 종목_거래일에 추가하여 (공통 거래집중도 ≥ 1) 필터를 적용한 계좌쌍들을 보여준다. 총 200개의 계좌쌍이 인지되었다. 이 계좌들은 자신의 상한가일 거래금액 전부를 상한가 종목 매매에 할당한 투자자들이다. 두

번째 행은 매수 거래금액관여율이 크고(공통 매수 거래금액관여율 ≥ 0.16), 세 번째 행은 호가공격성이 큰 (공통 호가공격성 ≥ 80 th percentile) 계좌쌍들을 추출하였다. 마지막 네 번째 행은 위의 지표들을 모두 다 적용한 결과를 보여주며, 그 결과 총 22개의 계좌쌍을 찾아내는 성과를 얻었다.

〈표 10〉 상한가 고정행위 혐의 연계계좌 추적을 위한 거래일과 연계지표의 결합

필터	계좌쌍수	상한가 종목_거래일	공통 거래집중도 최솟값	공통 매수거래 금액관여율 최솟값	공통 호가공격성 최솟값
(상한가 종목_거래일) and (공통 거래집중도 ≥ 1)	200	74	1		
(상한가 종목_거래일) and (공통 매수 거래금액관여율 ≥ 0.16)	353	91		0.16	
(상한가 종목_거래일) and (공통 호가공격성 ≥ 80 th percentile)	868	212			3.001
(상한가 종목_거래일) and (공통 거래집중도 ≥ 1) and (공통 호가공격성 ≥ 80 th percentile)	22	15	1		3.160

다음은 동일한 방법을 미공개정보 공개일에 적용하였다. 어떤 계좌가 기업의 미공개정보 공개일 이전에 활발히 거래하였다면 미공개정보를 이용할 가능성이 높다고 판단할 수 있다. 미공개정보 이용은 회사 임직원 등 내부자가 회사 기밀사항을 이용해 주식을 매매하거나 제자에게 알리는 행위이며, 자본시장법 제174조의 ‘미공개중요정보 이용행위 금지’ 조항에 따라 미공개 주요정보를 이용한 금융이익을 얻을 경우 형사처벌을 받게 된다. 한국거래소에서 다음과 같은 미공개정보이용 사례를 소개하고 있다. “甲과 乙은 각각 A사의 대표이사 및 상무로서 자기주식취득 검토를 지시한 후 동 사실을 공시하기 전에 동 정보를 이용하여 동사의 최대주주인 丙의 계좌를 통해 동사 주식을 매수하고, 차명계좌를 통해 동사 계산으로 자사주를 매수하여 부당이익을 취한 혐의가 있어 검찰고발 조치되었다.” 자사주취득 공시는 시장에서 호재성 공시로 인식을 하기 때문에 공시일 이후에 주가가 상승하는 것이 일반적이다. 미공개정보 이용자는 이를 미리 알고 공시일 전에 주식을 매수함으로써 부당이익을 취한 것이다.

사례에서 언급한 자사주취득 공시일 이 외에도 종목의 이익공시일, 인수합병의 발표일 등이 기업의 중요한 정보가 공시되는 날이다. 본 연구에서는 이 중 이익공시일을 사용하여 미공개보이용 혐의 연계계좌군을 적출해 보고자 한다. 분석을 위해 표본기간 2017년 3월부터 5월까지 기간에 존재하는 2017년 1분기 이익공시일을 조사하였을 때, 총 1,202개 종목의 공시일을 얻었다. 이 중에서 공시일 당일 수익률이 0보다 큰 공시일을 호재성 공시일이라고 정의하였으며, 총 461개의 종목의 공시일이 존재하였다. 이 호재성 공시일 전 (-5, -1) 거래일 사이에 집중 매수하는 형태를 보이는 계좌들을 미공개정보를 이용할 가능성이 큰 계좌라고 가정하였다. 호재성 공시일 전에 집중적으로 매수하는 계좌쌍들을 추출하기 위해 여러 연계지표들을 사용한 필터링을 실시하였다.

<표 11>은 그 결과를 보여주고 있다. 첫 번째 행은 호재성 이익공시일에 추가하여 거래집중도 필터(공통 거래집중도 ≥ 1)를 적용한 계좌쌍들을 보여준다. 총 33개의 계좌쌍이 인식되었다. 두 번째 행은 매수 거래금액관여율이 큰 계좌쌍 필터(공통 매수 거래금액관여율 $\geq 0.1^6$), 세 번째 행은 호가공격성이 큰 계좌쌍 필터(공통 호가공격성 $\geq 80\text{th percentile}$)를 적용한 경우를 보여준다. 마지막 네 번째 행은 위의 지표들을 동시에 적용한 결과를 보여주며, 그 결과 총 52개의 계좌쌍을 적출하였다.

<표 11> 이익공시일 미공개정보이용 혐의 연계계좌 추적을 위한 거래일과 연계지표의 결합

필터	계좌쌍수	종목_이익 공시일	공통 거래집중도 최솟값	공통 매수거래 금액관여율 최솟값	공통 호가공격성 최솟값
(이익공시 (-5, -1)거래일) and (공통 거래집중도 ≥ 1)	33	21	1		
(이익공시 (-5, -1)거래일) and (공통 매수 거래금액관여율 $\geq 0.1^6$)	37	18		0.16	
(이익공시 (-5, -1)거래일) and (공통 호가공격성 $\geq 80\text{th percentile}$)	69	22			3.038
(이익공시 (-5, -1)거래일) and (공통 거래집중도 ≥ 0.1) and (공통 호가공격성 ≥ 1)	52	15	0.113		1.182

5.3 3단계: 사회네트워크 분석 시행

앞에서와 같이 연계지표의 결합을 통해 또는 특정 사건일을 이용하여 혐의 연계계좌군을 찾아낼 수 있지만, 그 수가 많은 경우에는 더 범위를 좁혀야 할 필요성이 있다. 또한 계좌들 사이의 연결관계나 동적역학에 대해서는 파악하기 어렵다는 한계점도 존재한다. 이를 위해서 추가적으로 사회네트워크 분석(social network analysis)을 시도하고자 한다. 사회네트워크 분석은 계좌들이 실제적으로 어떻게 연결되어 있는지에 대해 더 구체적이며 다양한 정보를 제공해 줄 수 있다.

사회네트워크 분석을 통해 실제 계좌들 간의 네트워크를 그려볼 수 있다. 네트워크에서 계좌들이 어떤 요소에 의해서 연결되었는지에 따라서 다양한 관계를 표시할 수 있다. 이 변수를 가중치(weight)라고 하는데, 예를 들어, A와 B 계좌와의 관계를 특정지어 주는 다양한 변수들로 앞에서 구한 거래지표와 호가지표를 사용할 수 있다. 만약 가중치 변수로 공통 호가공격성 지표를 사용한다면, 이 지표가 클수록 더 강한 관계에 있는 계좌쌍으로 사회네트워크 분석에서 인식하게 된다. 한편 금융감독원 조사보고서에 의하면, 불공정거래의 평균 혐의 계좌수는 2015년 18개, 2016년 22개, 2017년 30개로 계속 증가하고 있는데, 사회네트워크 분석을 통해 이 정도 개체수가 포함된 연계 혐의계좌군을 찾아낸다면 현실과 비슷할 것이다.

다음은 Pump and Dump 불공정 행위를 예로 삼아 어떻게 사회네트워크 분석을 적용할 수 있는지 살펴보자. <표 7>의 마지막 행에 해당하는 105계좌쌍에 대해 사회네트워크 분석을

실시하고자 한다. 이들은 (공통 호가공격성 ≥ 80 th percentile), (공통 종목_거래일 ≥ 80 th percentile), (거래 집중도 ≥ 80 th percentile) 필터에 의해서 추출된 계좌쌍들이다. 계좌들 간의 관계에 대해서 공통 호가공격성 연계지표 값을 가중치로 사용하여 사회네트워크 분석을 실시하였다. 이들 간의 관계에 대해 <그림 2>의 (a)에서 보여준다. 점은 계좌번호를 나타내며, 연결선의 굵기는 가중치인 호가공격성에 의해서 결정된다. 호가공격성에서 밀접한 관계에 있는 계좌일수록 굵은 선으로 표현되고 있다. 105개의 계좌쌍에 해당되는 계좌들 간의 연결상태를 사회네트워크 분석을 통해 살펴볼 수 있다. 총 123개의 계좌들이 복잡하게 연결되어 있다. 네트워크의 평균 연결정도(average degree)는 1.707, 밀도(density)는 0.014이다.

핵심 하부구조를 찾기 위해, 추가적으로 K-core 분석을 실시하였다. Core의 개수를 2로 설정했을 때(2-core)의 결과를 아래 <그림 2>의 (b)에서 보여준다. 연결성이 약한 노드들은 제거되고 핵심이 되는 계좌 26개가 서로 밀접하게 연결되어 있어서 혐의 연계계좌군으로 인식가능하다. 특히 한 가운데 삼각형으로 가장 밀접하게 연계되어 있는 세 계좌 266688, 2218676, 2554284가 이 네트워크의 핵심이라고 할 수 있다. 이 세 계좌의 거래활동에 대해 더 세밀하게 조사할 수 있다면 더 많은 혐의 정보를 획득할 수 있을 것이다.

다음은 시가 및 종가 시세조종에 대해 적용해 보자. <표 9>에서 일단 시가(종가) 시세조종에 대한 혐의 연계계좌군을 찾기 위해 연계지표들을 결합하였을 때인데, 추출된 계좌쌍들의 수가 많지 않다. 이런 경우에 간단하게 사회네트워크를 그려봄으로써 여기에 포함된 계좌들 간의 관계를 파악할 수 있다. <그림 3>은 <표 9>의 첫째 행의 시가 시세조종 혐의 연계계좌군과 셋째 행의 종가 시세조종 혐의 연계계좌군에 대해 각각 보여주고 있다. 그림 (a) 시가 시세조종

<그림 2> Pump and Dump 혐의 연계계좌들의 네트워크

(a) 혐의계좌 전체

(b) 혐의계좌 컴퍼넌트(K-core 분석, k = 2)

혐의 계좌는 12개 계좌이다. 네트워크의 평균 연결정도는 1.333, 밀도는 0.121이다. 그림 (b) 종가 시세조종 혐의 계좌는 17개 계좌이다. 네트워크의 평균 연결정도는 1.333, 밀도는 0.078이다.

<그림 3> 시가 및 종가 시세조종 혐의 연계계좌들의 네트워크

세 번째, 상한가 고정 행위 사례로 A사 주식을 살펴보고자 한다. <그림 4>는 A사 주가를 보여주고 있다. 주가는 2017년 3~4월경에 최고가를 기록하였으며, 2017년 4월 5일 수익률이 20.28%로 상한가 수익률을 기록하였다. <표 10>은 2017년 4월 5일 A사 주식을 거래한 계좌들 중 연계지표에 의해 혐의계좌를 적출한 것이다. 분석을 위해, 세 번째 행의 필터링(공통 호가공격성 ≥ 80 th percentile)에 해당하는 계좌들 중에 A사 주식을 거래한 계좌쌍들을 추출하였다. 즉, 상한가일에 공격적인 매수 호가를 제출한 상위 계좌들을 인지하였다. 여기에 해당하는 계좌는 8개이며 이들은 서로 밀접하게 연결되어 있다.

<그림 4> A사 주가의 추이

<그림 5>의 (a)는 8개 계좌들의 관계를 공통 호가공격성 연계지표를 가중치로 사용하여 사회네트워크를 그린 것이다. 네트워크의 평균 연결정도는 6, 밀도는 0.857이다. 이전의 다른 네트워크에 비해서 노드들의 연결정도가 아주 강한 것을 볼 수 있다. 특히 상단에 있는 5개의 계좌가 밀접하게 연결되어 있는 것을 확인할 수 있다. <그림 5>의 (b)에서는 네트워크 상의 핵심 하부구조를 찾기 위해서 K-core 분석($k = 5$)을 실시하였다. 그 결과 여전히 상단에 있는

5개의 계좌가 밀접하게 연결되어 있는 것을 확인할 수 있다. 이 계좌들을 집중적으로 심리함으로써 더 많은 혐의점들을 찾아낼 수 있을 것이다.

<그림 5> 상한가 고정행위 혐의 연계계좌들의 네트워크: A사 사례

마지막, 이익공시일 미공개정보 이용 행위 사례로 D사 주식을 살펴보고자 한다. <그림 6>은 D사 주가를 보여주고 있다. 주가는 2017년 1분기 이익공시일인 2017년 4월 12일 최고가 3,365원을 기록하였다. 본 연구에서는 호재성 이익공시일 전 (-5, -1)거래일 동안 집중적으로 매수한 계좌들을 연계혐의 계좌로 가정하였다.

<그림 6> D사 주가의 추이

<표 11>은 D사의 2017년 1분기 이익공시일 5거래일 전 동안 해당주식을 활발히 거래한 계좌쌍을 적출해 보여주고 있다. 세 번째 행의 필터링(공통 호가공격성 $\geq$ 80th percentile)에 해당하는 계좌쌍들을 네트워크 분석에 사용하였다. 즉, 호재성 이익공시일 전 (-5, -1) 거래일 동안 공격적인 매수호가를 제출한 상위 계좌들을 인지하였다. 여기에 해당하는 계좌는 8개이며 이들은 서로 밀접하게 연결되어 있다.

<그림 7>의 (a)는 위의 8개 계좌들의 관계를 공통 호가공격성 지표를 가중치로 사용하여 사회네트워크를 그린 것이다. 네트워크의 평균 연결정도는 2.5, 밀도는 0.357이다. 특히 중앙에 있는 4개의 계좌가 밀접하게 연결되어 있는 것을 확인할 수 있다. <그림 7>의 (b)에서는 핵심 하부구조를 찾기 위해서 K-core 분석(k = 2)을 실시하였다. 그 결과 여전히 상단에 있는 4개의

계좌가 밀접하게 연결되어 있는 것을 확인할 수 있다. 이 계좌군들은 미공개정보 이용 혐의에 대해 강하게 의심할 수 있으며, 특히 중앙에 있는 2,712,177계좌는 네트워크의 중심에 있는 핵심 계좌이다. 이 계좌를 중심으로 집중 심리한다면 실제 연계계좌군을 찾는 데 큰 도움이 될 것이다.

〈그림 7〉 이익공시일 전 미공개정보이용 혐의 연계계좌들의 네트워크: D사 사례

지금까지 3단계의 과정을 통해서 계좌들 사이의 연계성을 계산하고 일정한 필터링을 통해서 혐의 연계계좌들을 찾아가는 과정을 제안하였다. 이는 여러 개의 수많은 계좌들 중에서 연계혐의 가능성이 낮은 계좌들을 버리고 높은 계좌들을 선택해가는 과정이라고 할 수 있다. 이런 소수 계좌들을 선택한다면 남은 계좌들의 혐의가능성을 높일 수 있을 것이다. 반면 이렇게 범위를 좁히는 과정 가운데서 실제 불공정거래 계좌를 놓쳐버릴 수 있는 반대급부도 존재할 수 있다. 하지만 이에 대해서는 크게 염려할 필요가 없다고 생각한다. 남아 있는 계좌들의 혐의 가능성을 높여서 실제 불공정거래를 행한 핵심계좌나 계좌군을 찾아 낼 수 있다면 이를 근거로 해서 다른 연계계좌들도 충분히 추적할 수 있기 때문이다.

6. 결론

본 연구는 단일계좌 중심의 시장감시에서 벗어나 연계계좌 중심의 불공정거래 탐색을 할 수 있는 방법론을 제안하는 것을 목적으로 하였다. 이를 위해 계좌 간의 연계정도를 측정할 수 있는 연계지표 20가지를 제안하고 계산하였다. 이후 이 연계지표를 사용한 사회네트워크 분석을 통해 계좌들 간의 연결정도와 중심성에 대해서 탐색하였다.

본 연구가 지니는 한계점과 향후 발전 방향에 대해 몇 가지 생각해 보았다. 첫째, 실제 연계계좌를 알 수 없는 한계 가운데 연구가 진행되었다. 연계지표를 사용하여 연계계좌쌍을 찾아낸다 하여도 이 계좌쌍이 실제 연계계좌인지는 알 수 없기 때문에, 본 연구에서 제안한 방법론의 성과를 측정할 수 없다. 금융감독원이나 한국거래소에서 실제 연계계좌쌍에 대한 정답 데이터를 제공해

줄 수 있다면 각 연계지표나 추출 방법들에 대한 평가가 더 정확하게 이루어 질 수 있을 것이다. 둘째, 연계계좌 적출과정 중에 더 연구하고 최적화시켜야 할 부분들이 있다. 본 연구에서는 거래지표 10개, 호가지표 10개를 제안하였지만 사용자의 목표에 따라서 더 다양한 지표를 개발하여 사용할 수 있다. 또한 연계지표를 사용한 필터링에 있어서 각 지표의 절단점(cutoff point)에 대해서도 더 심도있는 연구가 필요하다. 즉, 각 지표에 대해서 어떤 절단점을 사용하여 혐의 연계계좌들을 필터링할 때 가장 효율적으로 연계계좌를 특정할 수 있는지에 대해 본 연구는 분석하지 못하였다. 이들에 대해서는 향후 더 연구되길 기대해 본다.

참 고 문 헌

- 김상락, 강만모, “빅데이터 분석 기술의 오늘과 미래”, 정보과학회지, 제32권 제1호(2014), pp. 8-17.
- 김석수, “금융 상품 추천에 관련된 빅 데이터 활용을 위한 개발 방법”, 한국컴퓨터정보학회논문지, 제19권 제8호(2014), pp. 73-81.
- 김재생, “빅데이터 분석 기술과 활용사례”, 한국콘텐츠학회지, 제12권 제1호(2014), pp. 14-20.
- 김정숙, “빅 데이터 활용과 관련기술 고찰”, 한국콘텐츠학회지, 제10권 제1(2012), pp. 34-40.
- 박연길, 안희준, 여환영, “종가시세조종이 주가와 거래량에 미치는 영향”, 한국증권학회지, 제42권 1호(2013), pp. 285-321.
- 양철원, “선물·옵션 만기일의 주가조작 행위에 대한 실증분석”, 한국증권학회지, 제47권 5호 (2018), pp. 709-739.
- 양철원, 유지연, “만기일 효과에 지수차익거래의 영향: 정상거래인가? 연계시세조종인가?”, 재무연구, 제29권 제2호(2016), pp. 149-191.
- 양철원, 유지연, “연계시세조종 행위에 대한 규제: 도이치은행 사례를 중심으로”, 한국증권학회지, 제46권 1호(2017), pp. 217-248.
- Barclay, M. J. and J. B. Warner, 1993, Stealth trading and volatility: Which trades move prices?, *Journal of Financial Economics* 34 (3), pp. 281-305.
- Bernhardt, D. and R. J. Davies, 2005, Painting the tape: Aggregate evidence, *Economics Letters* 89 (3), pp. 306-311.
- Carhart, M. M., R. Kaniel, D. K. Musto, and A. V. Reed, 2002, Leaning for the tape: Evidence of gaming behavior in equity mutual funds, *Journal of Finance* 57 (2), 661-693.
- Comerton-Forde, C. and T. J. Putnins, 2011, Measuring closing price manipulation, *Journal of Financial Intermediation* 20 (2), pp. 135-158.
- Hillion, P. and M. Suominen, 2004, The manipulation of closing prices, *Journal of Financial Markets* 7 (4), pp. 351-375.
- Khwaja, A. I. and A. Mian, 2005, Unchecked intermediaries: Price manipulation in an emerging stock market, *Journal of Financial Economics* 78 (1), pp. 203-241.
- Kumar, P. and D. J. Seppi, 1992, Futures Manipulation with “Cash Settlement”, *Journal of Finance* 47 (4), pp. 1485-1502.

A Study on Detecting the Linked Accounts by Using Big-Data Analysis^{*}

Cheol-Won Yang^{}**

Associate Professor, Dankook University

Abstract

The most significant feature of recent unfair trade transactions is the increase in the number of suspicious accounts. Despite the fact that the number of unfair transactions investigated by the FSS has decreased every year, the average number of suspicious accounts continues to increase. This means that many unfair trading practices are conducted through multiple linked accounts rather than single accounts. In this study, I develop a linkage index for a pair of accounts, and propose a method of extracting the linked accounts and databaseing them using social network analysis. Since the analysis of the linked account requires a lot of calculation time, the market surveillance organizations can recognize the linked accounts in advance and investigate it when there are abnormal signs in the transaction. In addition, after processing the actual unfair transaction event, they can update the linked account to build a more sophisticated linked account database. This process will further strengthen market surveillance on unfair trade practices.

Keywords: Unfair Trade; Linked Accounts; Linkage Index; Social Network Analysis; Market Surveillance

JEL Classification: G14, G18, G28

* This research was supported by the Soundness of Stock Market Grant Program from the Korea Exchange and Korean Securities Association in 2018.

** Corresponding Author. Address: School of Business Administration, Dankook University, 152 Jukjeon-dong Suji-gu Yongin-si Geonggi-do, Korea, 16890; E-mail: yang@dankook.ac.kr; Tel: +82-31-8005-3437; Fax: +82-31-8021-7208.