# Insider Trading Enforcement & Link Prediction*

## Introduction

The Securities and Exchange Commission (SEC) has taken a very aggressive stance towards insider trading enforcement, promising to be an unrelenting and omnipresent foe to those who seek to abuse nonpublic information at the expense of investors. But in reality, the SEC's resources are limited. Until recently, the SEC was forced to rely on outside tips as its primary means of detection and to triage its cases heavily—pursuing only the lowest hanging fruit. Today, however, technology provides a means for the SEC to get more bang for its buck. To date, the SEC's tech-enforcement efforts have been centered on the development of in-house, automated, market-data-analysis systems to proactively detect suspicious trades, thereby reducing its reliance on outside tips. But despite the increasing ease with which investigators are able to pinpoint illegal trades amidst a din of market data, in cases where no readily discernable insider connection exists, the SEC must continue its old triage practices, mothballing potentially viable cases due to expected investigation costs. In short, despite its new toys, the SEC continues to suffer one of its old bottlenecks.

To that end, link prediction and corollary algorithmic technologies may be just the shot in the arm the SEC needs to better manage its caseload. Link prediction works by using network structures and other information to predict connections that should or will exist—for example, between people or organizations. In theory, with enough information, link prediction systems could predict traders' connections to inside sources. Investigators may be able to use these systems as a means to conduct low-cost preliminary case audits and, should a link be predicted, build a roadmap for a full investigation.

But despite the allure of this powerful tool, its application faces practical and legal hurdles. First, as a practical matter, link prediction systems are limited in their ability to synthesize dissimilar data into a usable format and in their ability to filter out false-positives (i.e., irrelevant connections). Fortunately for the SEC, ongoing research into data synthesis and link-strength prediction is helping to mitigate these issues, and link prediction technology is not so limited that it cannot be used in its current form. Second, as a legal matter, it may be difficult to satisfy the Fourth Amendment when obtaining certain data, to square algorithmic enforcement with notions of due process, and to rely on statistical data to sustain a verdict. Nevertheless, thanks to the third-party doctrine and the Stored Communications Act, the

* I thank David Garrett for guiding me through this spelunk into network analysis.

SEC will have almost free access to a wealth of transactional data—the type of data most useful in link prediction. Thus, at a bare minimum, link prediction systems can serve as a tool to inexpensively vet cases and point investigators in the right direction, ameliorating the current triage bottleneck.

This Note briefly reviews insider trading liability before moving into the SEC's role vis-à-vis insider trading, the issues it faces, and its current tech-enforcement efforts. From there, this Note discusses briefly the basics of link prediction, data synthesis, and link-strength prediction, and then turns to the application of those technologies and the limits imposed on them by the Constitution. Finally, this Note concludes with a worst-case scenario for the SEC should it decide to incorporate link prediction analysis into its current enforcement scheme.

## I. An Overview of Insider Trading Liability

Many are familiar with the concept of insider trading: an "insider" with information unknown to the public makes a trade in order to profit from the market's ignorance. However, not all trades based on nonpublic information are illegal. Illegal insider trading, which is referred to here simply as "insider trading," requires, essentially, the "buying or selling [of] a security, in breach of a fiduciary duty or other relationship of trust and confidence, while in possession of material, nonpublic information about the security."[1] In short, (1) did the trader trade on nonpublic information, and (2) did the trader have a duty to disclose the information before doing so? In the second question lies the rub.

17 C.F.R. § 240.10b–5 (Rule 10b–5),[2] which contains more pointed (albeit still ambiguous) antifraud provisions than the statute under which it was promulgated,[3] has made it "relatively easy" to prohibit company *insiders* from profiting off of nonpublic information because "such behavior fit[s]

1. *Fast Answers: Insider Trading*, U.S. SEC. & EXCHANGE COMMISSION (Jan. 15, 2013), https://www.sec.gov/fast-answers/answersinsiderhtm.html [https://perma.cc/CP9Q-ULYX].

2. 17 C.F.R. § 240.10b–5 (2017) reads:
> It shall be unlawful for any person, directly or indirectly, by the use of any means or instrumentality of interstate commerce, or of the mails or of any facility of any national securities exchange, (a) To employ any device, scheme, or artifice to defraud, (b) To make any untrue statement of a material fact or to omit to state a material fact necessary in order to make the statements made, in the light of the circumstances under which they were made, not misleading, or (c) To engage in any act, practice, or course of business which operates or would operate as a fraud or deceit upon any person, in connection with the purchase or sale of any security.

3. 15 U.S.C. § 78j (2016) forbids traders
> to use or employ, in connection with [the sale of securities,] . . . any manipulative or deceptive device or contrivance in contravention of such rules and regulations as the [Securities and Exchange] Commission may prescribe as necessary or appropriate in the public interest or for the protection of investors.

within traditional notions of fraud."[4] But trades based upon nonpublic information made by *noninsiders* too can bear the markings of fraud. To that end, courts have used Rule 10b–5 as a versatile host, onto which they engrafted two noninsider-insider trading doctrines: tippee and misappropriation liability. The two defining cases for tippee and misappropriation liability are *Dirks v. SEC*[5] and *United States v. O'Hagan*[6] respectively. In *Dirks*, the Supreme Court held that to hold a tippee liable for trading on nonpublic information, the tipper must "breach[] [a] fiduciary duty to the shareholders by disclosing the information to the tippee," and the tippee must "know[] or should know that there has been a breach."[7] Fourteen years later, in *O'Hagan*, the Supreme Court held that traders who normally owe no fiduciary duty not to trade on nonpublic information can be liable when they acquire nonpublic information through the "deception of those who entrusted [them] with access to [that] information."[8]

The takeaway here is that one does not have to be an employee of a company to be liable for insider trading. Liability depends on how one obtained nonpublic information. An insider trader could be a contractor, a tennis buddy, a distant relative, or a friend of a friend. Needless to say, the task of tracking down a trader's inside connection can be an arduous one.

## II. The SEC's Enforcement Division and Its Use of Technology

The SEC is not only tasked with promulgating rules under the 1934 Act but also with the Act's enforcement. Accordingly, a driving force against insider trading is the SEC's Enforcement Division, which conducts detection efforts and investigations, and handles enforcement proceedings on behalf of the SEC.[9] Before turning to the SEC's tech-enforcement practices, it is

4. Thomas C. Newkirk, Assoc. Dir., Div. Enf't, U.S. Sec. & Exch. Comm'n, & Melissa A. Robertson, Senior Counsel, Div. Enf't, SEC, Insider Trading—A U.S. Perspective, Remarks at the 16th International Symposium on Economic Crime, Jesus College, Cambridge, England (Sept. 19, 1998), https://www.sec.gov/news/speech/speecharchive/1998/spch221.htm [https://perma.cc/492C-FJEN].

5. 463 U.S. 646 (1983).

6. 521 U.S. 642 (1997).

7. *Dirks*, 463 U.S. at 660. Although the Court did not explain every way in which one might breach one's duty through tipping, the Court did discuss one particular manner: receiving a benefit in return for the tip. *Id.* at 662. The Court noted that there is an assumption of benefit where there is "an intention to benefit the particular recipient." *Id.* at 664. This is because a gift of inside information for the purposes of trading is functionally equivalent to committing insider trading oneself then giving the proceeds to the recipient. *Id.* This theory of breach was recently reinforced in *Salman v. United States*, 137 S. Ct. 420, 422–24 (2016).

8. *O'Hagan*, 521 U.S. at 652.

9. Chien-Chung Lin & Eric Hung, *U.S. Insider Trading Law Enforcement: Issues and Survey of SEC Actions from 2009 to 2013*, 11 NAT'L TAIWAN L. REV. 37, 46 (2016). Insider trading cases constitute nearly 8% of all yearly SEC enforcement actions. *Id.* at 65.

important to start with the basics of SEC enforcement and the SEC's philosophy on insider trading.

### *A. SEC Enforcement*

As a general matter, enforcement actions can be divided into four stages: detection, preliminary investigation, formal investigation, and prosecution.

*1. Detection.*—The Enforcement Division gathers information regarding potential violations from public, private, and internal sources. Common sources include periodic filings; complaints from investors, whistleblowers, and competitors; the media; referrals from self-regulatory organizations like the Financial Industry Regulatory Authority (FINRA) (which total in the hundreds yearly);[10] and in-house market surveillance, which is discussed at length below.[11]

*2. Preliminary Investigation.*—The Division sifts through its repository of nascent cases to determine which matters should be designated Matters Under Inquiry (MUI).[12] MUIs receive an informal investigation. During an informal investigation, investigators attempt to develop the facts "through . . . interviewing witnesses, examining brokerage records, reviewing trading data, and other methods."[13] Following an informal

10. *See* Christopher P. Montagano, *The Global Crackdown on Insider Trading: A Silver Lining to the "Great Recession,"* 19 IND. J. GLOBAL LEGAL STUD. 575, 595 (2012) (stating that "[i]n 2010, FINRA referred 244 insider trading cases to the SEC, the highest in the history of FINRA").

11. Lin & Hung, *supra* note 9, at 48.

12. Because the Division cannot follow up on every matter brought to its attention, it must triage its investigations. The Division uses MUI designation as a threshold filter "to help ensure efficient allocation of resources." SEC. & EXCH. COMM'N DIV. ENF'T, ENFORCEMENT MANUAL § 2.3.1 (Oct. 28, 2016), https://www.sec.gov/divisions/enforce/enforcementmanual.pdf [https://perma.cc/7C23-BZPX]. MUI analysis has only two prongs: "whether the facts underlying the MUI show that there is potential to address conduct that violates the federal securities laws" and "whether the assignment of a MUI to a particular office will be the best use of resources for the Division as a whole." *Id.* Naturally, this test favors those cases that appear to be slam dunks *ex ante* and disfavors those that do not—regardless of whether they are in fact viable. Complementary to MUI designation, Division staff also rank open investigations for further resource allocation, dubbing certain cases "National Priority Matters." *Id.* § 2.1.1. NPM status is reserved for those cases that offer an "opportunity to send a particularly strong and effective message of deterrence," involve "egregious or extensive misconduct" that threatens "widespread and extensive harm to investors," and feature perpetrators that "occupy[] positions of substantial authority or responsibility." *Id.*

13. *How Investigations Work*, U.S. SEC. & EXCHANGE COMMISSION (Jan. 27, 2017), https://www.sec.gov/enforce/how-investigations-work.html [https://perma.cc/RG6U-SUXR].

investigation, the Division decides whether to seek a formal order of investigation from the Commission.[14]

*3. Formal Investigation.*—Upon obtaining a formal order of investigation, the Division may compel witnesses and demand the production of documents.[15] The primary focus of the investigation is to build the case. One internal memo states that investigators are to focus on the specific elements of the case: possession of nonpublic information, materiality, scienter, and duty.[16] After the investigation is complete, the Commission reviews the findings and decides how to proceed.[17] It may go forward by filing in Federal Court or initiating an administrative proceeding.[18] It may also decide to refer the case to the Department of Justice for parallel criminal proceedings.[19]

*4. Prosecution.*—The Division prosecutes cases in the name of the SEC.[20]

### *B. The SEC's Philosophy*

According to the SEC, rooting out insider trading is critical to maintaining the "highest level of confidence" in the market, which gives investors the confidence to "put their fortunes at risk."[21] That said, the SEC faces a wealth of potential investigations while operating with finite resources.

---

14. *Id.*

15. 15 U.S.C. § 78u(b) (2012).

16. L. HILTON FOSTER, U.S. SEC. & EXCH. COMM'N., INSIDER TRADING INVESTIGATIONS 4–5 (Sept. 2000), https://www.sec.gov/about/offices/oia/oia_enforce/foster.pdf [https://perma.cc/MXR2-B6KK].

17. *How Investigations Work*, *supra* note 13.

18. *Id.*

19. Lin & Hung, *supra* note 9, at 48.

20. *Id.* at 46–47.

21. Arthur Levitt, Chairman, U.S. Sec. & Exch. Comm'n, A Question of Investor Integrity: Promoting Investor Confidence by Fighting Insider Trading, Address Before the "SEC Speaks" Conference (Feb. 27, 1998), https://www.sec.gov/news/speech/speecharchive/1998/spch202.txt [https://perma.cc/ZD5A-ZUMC].

*1. No Place to Hide.*—Former SEC Chair Mary Jo White once urged that the SEC must play the part of the bold and unrelenting enforcer, "perceived to be . . . everywhere."[22] She continued,

> [I]nvestors . . . want to know that there is a strong cop on the beat—not just someone sitting in the station house waiting for a call, but patrolling the streets and checking on things.
>
> They want to know that would-be fraudsters are spending more time looking over their shoulders, and less time stepping over the line.[23]

Although White acknowledged the SEC does its best "not to disappoint," she also acknowledged that omnipresence is more difficult to achieve in "today's fast moving, complex and changing markets."[24]

*2. No Case Too Small.*—To engender investor confidence, the SEC seeks to "pursue all wrongdoers—individual and institutional, of whatever position or size."[25] "Investors," White has said, "do not want someone who ignores minor violations, and waits for the big one that brings media attention."[26] This no-case-too-small mentality has also been championed by former Chair Arthur Levitt, who said, "It's not as if insider traders wander innocently into gray areas near the boundaries of legality. They willfully stride across the bright line of the law."[27] Nevertheless, other SEC sources disclose that, "given the inherent difficulties in investigating and proving insider trading cases, the reality is that there is a significant amount of clearly illegal activity that goes undetected or unpunished."[28]

*3. No Resources Too Finite?*—The SEC, like all law enforcement agencies, is an entity of finite resources. In the words of Chair White, enforcement resources are "not nearly sufficient to the enormity and scope of the responsibility [the SEC has]."[29] As discussed below, this has created

---

22. Mary Jo White, Chair, U.S. Sec. & Exch. Comm'n, Remarks at the Securities Enforcement Forum (Oct. 9, 2013), https://www.sec.gov/news/speech/spch100913mjw [https://perma.cc/JU3Q-6RH3].

23. *Id.*

24. *Id.*

25. Michael D. Trager et al., *Mary Jo White Confirmed to Chair SEC; "Bold" Enforcement Envisioned*, ARNOLD & PORTER LLP 2 (Apr. 2013), http://files.arnoldporter.com/adv514recentdevelopmentsinsecenforcementreviewofwhite.pdf [https://perma.cc/G9SV-CN5G].

26. White, *supra* note 22.

27. Levitt, *supra* note 21.

28. Newkirk & Robertson, *supra* note 4.

29. White, *supra* note 22 ("In the first instance, we of course recognize that our resources are not infinite.").

tremendous tension between the SEC's need for omnipresence and unrelenting enforcement, and its checkbook.

### *C. Traditional Enforcement Methods*

Traditionally, the Enforcement Division (the Division) has taken a "security-based" approach to insider trading detection.[30] Under this approach, the Division either receives a tip about potential abuses or learns of a disclosure of material nonpublic information.[31] It then inquires into who fortuitously traded the security in question and their motivation for doing so.[32] Until recently, this was the go-to detection approach because detecting illicit trades as they occurred was extremely difficult.[33] Thus, despite being reactive and inefficient, it was the only practical means.[34]

When suspicious trades are detected, unless a connection to an insider is readily apparent, "one of the most challenging issues [an investigation faces] is establishing the relationship [to an insider]."[35] To do so, the Division has to beat the proverbial bushes and interview witnesses and review documents in the hope of flushing out a link.[36] Even wiretapping, publicly showcased in the Rajaratnam trial as a powerful tool for investigating insider trading,[37] is exceptionally labor-intensive for investigators, both in terms of obtaining authorization from the court and in terms of running the tap operation itself.[38] The resource-intensive nature of traditional investigation methods forced the Division to adopt a policy of assessing the viability of

30. Jake Steele, *SEC Utilizes Big Data to Fight Insider Trading*, CONSUMERS' RES. (Nov. 1, 2016), http://consumersresearch.org/sec-utilizes-big-data-to-fight-insider-trading/ [https://perma.cc/QE5J-GTU6].

31. *Id.*

32. *Id.*

33. *Id.* Only large and extremely fortuitous trades would jump out at anyone reviewing market data. Lin & Hung, *supra* note 9, at 53.

34. Steele, *supra* note 30.

35. White, *supra* note 22.

36. For a list of the multitude of witnesses and documents useful in an insider trading investigation, see generally the Foster memo, *supra* note 16.

37. For more information on the Rajaratnam wiretapping, see Eyder Peralta, *The Wiretaps that Built the Case Against Galleon's Rajaratnam*, NPR: THE TWO-WAY (May 11, 2011), http://www.npr.org/sections/thetwo-way/2011/05/11/136206203/the-wiretaps-that-built-the-case-against-galleons-rajaratnam [https://perma.cc/H9Q6-WC33].

38. 18 U.S.C. § 2518(1) (2012) provides that investigators must file an application "in writing . . . to a judge of competent jurisdiction" and include, among other things,

> details as to the particular offense[,] . . . a particular description of the nature and location of the facilities from which . . . the communication is to be intercepted, . . . a particular description of the type of communications sought[,] . . . the identity of the person, if known, . . . whose communications are to be intercepted[,] . . . [a] statement as to whether or not other investigative procedures have been tried . . . or why they reasonably appear to be unlikely to succeed, . . . [and] a statement of the period of time for which the interception is required to be maintained.

cases *ex ante* based upon what little information was initially available, thereby ensuring a lowest-hanging-fruit method of enforcement.[39]

### *D. The Move to Tech-Enforcement*

Because the Division's resources are limited, it increasingly turns to technology to realize new efficiencies. To date, SEC efforts have been focused primarily on building in-house market-analysis systems to detect suspicious trades.[40] This new mode of detection has been dubbed the "trader-based" approach to enforcement.[41] In recent years, the Division's Market Abuse Unit has unveiled three market-analysis programs, which together form an impressive toolbox/alphabet soup: ARTEMIS,[42] ABAP,[43] and NEAT.[44] These programs are designed to detect both suspiciously fortuitous trades and suspiciously coordinated trades.[45] This approach not only provides a comparatively proactive means of detection, it also provides a means of detecting insider trading rings.[46] In short, the SEC's new technological tools provide an inexpensive way to self-tip and, in the case of coordinated trades, inexpensively tie conspirators together. This has allowed the SEC to do some of "[w]hat used to take weeks . . . in a matter of hours."[47]

---

39. *See supra* note 12 and accompanying text.

40. Trager et al., *supra* note 25, at 4.

41. Steele, *supra* note 30.

42. *Id.* (describing the Advanced Relational Trading Enforcement Metrics Investigation System, which is designed to detect suspicious trading patterns between traders over time).

43. White, *supra* note 22 (describing the Advanced Bluesheet Analysis Program, which is designed to analyze specific securities transactions to detect "suspicious trading before market-moving events" and flag coordinated trades).

44. Mary Jo White, Chair, Sec. & Exch. Comm'n, Opening Remarks at the 21st Annual International Institute for Securities Enforcement and Market Oversight (Nov. 2, 2015), https://www.sec.gov/news/statement/remarks-21st-international-institute-for-securities-enforcement.html [https://perma.cc/8AUE-9B5M] (describing the National Exam Analytics Tool, which is designed to allow investigators to rapidly access transactions from a massive index).

45. Steele, *supra* note 30. Studies testing the accuracy of these detection methods have so far been positive. *See, e.g.*, Acar Tamersoy et al., *Inside Insider Trading: Patterns & Discoveries from a Large Scale Exploratory Analysis*, GA. TECH C. COMPUTING (Aug. 25–28, 2013), http://www.cc.gatech.edu/~dchau/insider/asonam13_insider.pdf [https://perma.cc/8B39-Q324] (describing synced-transaction analytics as "a major step" in detecting insider trading networks). Interestingly, private companies have also taken to using this technology to detect their own leaks. *See, e.g.*, *Insider Trading Detection*, FINANCIAL TRACKING, http://www.financial-tracking.com/insider-trading-detection/ [https://perma.cc/HX4P-BZJG] (describing a detection system that "alert[s] if a trade occurred prior to a market moving issuer event").

46. Steele, *supra* note 30. Recently, "suspicious trading patterns" led the SEC to charge 34 people in connection to an insider trading scheme. White, *supra* note 22.

47. White, *supra* note 22.

But until this year, the SEC did not have a direct feed to the various stock exchanges.[48] It pulled data on an ad hoc basis and cobbled together its own data pool for processing.[49] Despite the database containing "over 6 billion electronic equities and options trading records,"[50] this ad hoc process meant an unrealized analytics capability as the lack of complete market data left gaps in the record and skewed the data pool. As a solution, the SEC recently announced it would begin aggregating all equity- and option-market trading activity through a centralized system it calls the Consolidated Audit Trail (CAT).[51] The CAT system will compile every trade order, execution, and cancellation for processing, pulling both from national stock exchanges and FINRA databases.[52] By enabling the SEC to use its algorithmic detection systems on all national-market trading data, the CAT system stands to advance the SEC one step closer towards its goal of appearing to be everywhere at once.

### *E. So, Where Does the SEC Go from Here?*

By honing its ability to detect illegal trades and the coordinated efforts of insider trading rings, the SEC frees detection and investigation resources. This, in turn, allows for more investigations, thereby reducing the load on its triage system. But for those cases where links are not readily apparent, the Division continues to face "one of the most challenging issues" in enforcement.[53] Accordingly, while the SEC has made important steps towards achieving omnipresence and resource efficiency, it continues to suffer from its bottleneck at the investigation stage.[54]

But the bottleneck may not exist for long. Advances in link prediction technology may offer a way to predict connections to insiders by relying on traders' social network data and similar information. If the SEC can incorporate this technology, it should conserve investigation resources, allowing for additional investigations. Moreover, it should redeem a number of cases that would otherwise be left on the cutting room floor in the current triage system. But before addressing the practical capabilities and legal ramifications of using link prediction technology, in the name of plenitude, the basics of link prediction and its corollaries must be explained.

---

48. Nate Raymond, *Newest Weapon in U.S. Hunt for Insider Traders Paying Off*, REUTERS (Nov. 1, 2016), http://www.reuters.com/article/usa-insidertrading/rpt-insight-newest-weapon-in-u-s-hunt-for-insider-traders-paying-off-idUSL1N1D200U [https://perma.cc/RRW4-BUWB].

49. *Id.*

50. White, *supra* note 44 (referring to establishing connections to insiders).

51. *Rule 613 (Consolidated Audit Trail)*, U.S. SEC. & EXCHANGE COMMISSION, https://www.sec.gov/divisions/marketreg/rule613-info.htm [https://perma.cc/USL8-28TZ].

52. *Id.*

53. White, *supra* note 22.

54. *See supra* note 12 and accompanying text.

## III. Detecting Links: Link Prediction, Data Synthesis, and Link-Strength Prediction

### *A. Introduction to Network Analytics*

Imagine your social network. It consists of connections to friends, family, and acquaintances. At only two degrees of separation, you are connected to every friend, family member, and acquaintance of your friends, family members, and acquaintances. Taking this to its logical conclusion, if one were not limited by degrees of separation, it would be odd not to find a connection between any two people on Earth. When one considers that meaningful connections may exist between people not only through social relationships but also through other connections—for example, two strangers working for the same company—the paths seem infinite.

Algorithms are adept at tracing the shortest connection between two points on a network using breadth-first bilateral analysis.[55] With enough information, a program could detect a trader's link to an insider with little difficulty. Unfortunately for the SEC, information does not come freely, nor is it guaranteed to come in a form that will mesh easily for analysis. Moreover, even if the veil can be lifted enough to find a link, it may not be the correct link. Nevertheless, these barriers have recently started to crumble as researchers in link prediction, data synthesis, and link-strength prediction continue to hone a capable network-analysis toolkit that is well-suited to insider trading investigation.

### *B. Link Prediction*

*1. What Is Link Prediction?*—Link prediction, in its most formal sense, refers to predictions as to where links will be made in the future. However, the term is also commonly used to refer to existing-link prediction—sometimes called inferring missing links.[56] For the purpose of this Note, I

---

55. "Breadth first" refers to a search pattern that establishes all branches at one degree of separation before moving to the next degree. Prateek Garg, *Breadth First Search*, HACKEREARTH, https://www.hackerearth.com/practice/algorithms/graphs/breadth-first-search/tutorial/ [https://perma.cc/P2W2-32MY].

56. *See, e.g.*, Michael Fire et al., *Computationally Efficient Link Prediction in a Variety of Social Networks*, 5 ACM TRANSACTIONS INTELLIGENT SYS. & TECH. 10:1, 10:1 (2013) (using the terms interchangeably). Fortunately for the SEC, attempts to predict existent links based upon present data have fared far better than attempts to predict future links based upon present data. This is in part because existent link prediction relies on a mix of structural and nonstructural information, rather than only on information regarding structural evolution, and in part because as time progresses, actual links develop linearly while negative links (links that do not occur) grow quadratically. Ahmad Sadraei, *Link Prediction Algorithms: What Will Facebook Friendships Look Like Tomorrow?* (Spring 2014), http://be.amazd.com/link-prediction/ [https://perma.cc/KR5T-YZL4]; *see also* David Liben-Nowell & Jon Kleinberg, *The Link-Prediction Problem for Social*

refer to existing-link prediction simply as link prediction, the form relevant to an insider trading investigation.

*2. The Context in Which Link Prediction Is Normally Discussed.*—Link prediction is often discussed in the context of social media, ostensibly because social media provides researchers with ample information with which to experiment.[57] Indeed, being able to better recommend friends on Facebook would be a very obvious use of link prediction algorithms. However, no link prediction study worth its salt misses the opportunity to propose a novel application for link prediction. Notable examples include filling in information gaps in web-crawling software sweeps,[58] finding interactions between proteins,[59] and identifying members of terrorist organizations.[60]

Although link prediction papers often note a potential application in criminal network detection, there has been little in the way of actual studies.[61] Among those studies that have discussed the use of link prediction in criminal investigations, its application to insider trading has only been mused about in passing.[62] This is regrettable because unlike the criminal-investigation applications that garner attention—namely, filling out the margins of larger conspiracies or detecting terrorist networks—insider trading investigations

*Networks*, 58 J. AM. SOC'Y INFO. SCI. & TECH. 1019, 1030 (2007) (obtaining only 16% accuracy predicting future links).

57. For example, one study ran tests using the members of pet-enthusiast websites Dogster, Catster, and Hamsterster. *See* Elena Zheleva et al., *Using Friendship Ties and Family Circles for Link Prediction*) ("On these sites, profiles include photos, personal information, characteristics, as well as membership in community groups. Members also maintain links to friends and family members."), *in* ADVANCES IN SOCIAL NETWORK MINING AND ANALYSIS 97, 103 (Lee Giles, Marc Smith, John Yen & Haizheng Zhang eds., 2010).

58. Fire et al., *supra* note 56, at 10:2.

59. *Id.*

60. Emrah Budur et al., *Structural Analysis of Criminal Network and Predicting Hidden Links Using Machine Learning*, ARXIV (Sept. 21, 2015), https://arxiv.org/pdf/1507.05739.pdf [https://perma.cc/PGJ6-GPN3].

61. Giulia Berlusconi et al., *Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis*, 11 PLOS ONE, no. 4, 2016, at 1, 2–3, http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0154244&type=printable [https://perma.cc/3A56-2GQD] ("This is surprising, not only given the significant growth of works on missing links in other fields [and] with the development of a number of different strategies, but also given that criminal investigations face the problem of missing links almost by definition, due to the scarcity of investigative resources and the antidetection strategies by criminals." (footnotes omitted)).

62. *See* Andrew Arnold & William W. Cohen, Information Extraction as Link Prediction: Using Curated Citation Networks to Improve Gene Detection 2 (Jan. 1, 2009) (unpublished manuscript) (on file with the National Center for Biotechnology Information), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018763/pdf/nihms191712.pdf [https://perma.cc/4G6N-H8M2] (noting a potential application in detecting "insider trading cabals").

are particularly poised to benefit from link prediction.[63] As discussed above, these investigations increasingly begin with nothing more than a suspicious trade. Consequently, the case will rise or fall on whether a connection to inside information can be established or, just as important in the case of the SEC's triage system, how much effort will be required to establish a connection.

*3. How It Works.*—This Note will not attempt to explain the intricacies of link prediction.[64] However, a basic understanding of link prediction is necessary to understand how to apply link prediction to an insider trading investigation. At the highest level of abstraction, link prediction systems work by predicting the existence or nonexistence of links based upon the existence of "features" that are similar to those they have seen before.[65]

Network features that speak to the network structure itself are called topographical features.[66] We almost always envision a network as a topographical model of the connections it contains. A topographical model can conceptualize a few different features. First, the topographical features can be comprised of dyadic relationships (e.g., person-to-person links with each person representing a "node") or of spheres of group membership, or a combination of the two.[67] Looking for commonality between two nodes based on the topographical features of their individual networks is often referred to as a test of "structural equivalence." As a rudimentary example of structural equivalence, consider that criminal networks are often arranged into a hierarchy.[68] It might surprise no one to find that two known criminals who live in the same city and whose social-network topographies contain

63. Berlusconi et al., *supra* note 61, at 3 (noting that the role of link prediction in law enforcement is not only a means of detecting the otherwise undetectable, but also a means to conserve "scarce investigative resources").

64. I am as relieved as you are.

65. Machine learning has become the prevailing methodology for link prediction. Machine learning refers to systems that are designed to generalize from the information they have previously processed to make inferences based upon new data. *E.g.*, Fire et al., *supra* note 56, at 10:2 (training machine-learning classifiers "on a set of easy-to-compute topological features"); *cf. id.* (using backward-looking Bayesian methods as an alternative or complement to supervised machine learning). One machine-learning study suggests that for social-network prediction, "near-maximal [area under the curve (AUC)] can be obtained by using training sets with up to 50,000 examples." *Id.* at 10:13. AUC under the receiver operating characteristic (ROC) curve is the "de facto performance measure for link prediction tasks." *Id.* at 10:6.

66. For example, the number of people one knows is considered a topographical feature.

67. Zheleva et al., *supra* note 57, at 99.

68. Budur et al., *supra* note 60, at 2.

traces of a similar hierarchical structure are part of the same criminal organization.[69]

But beyond topography, link predictors also consider descriptive features: those without regard to network structure.[70] A rudimentary example of a descriptive feature might be the dog one prefers. Again, it might surprise no one that two bichon frise enthusiasts are both connected to each other through a bichon frise-centric social group.[71]

*4. Study Results.*—Despite some studies' very impressive results, it is not called the "link-prediction problem" for no reason. While existent link prediction at one degree of separation between nodes has been shown to be a simple feat,[72] accurately predicting multi-degree links (e.g., based on X's data, X should know Y, and Y should know Z) is far more difficult, although still possible.[73]

### *C. Data Synthesis*

Today, we store an incredible amount of electronic data, and our stockpile grows exponentially. By one account it doubles in volume every two years at the current rate of five quintillion bytes produced every two days.[74] Practically hemorrhaging data in our daily lives, information from our phone calls, credit card purchases, and internet usage (emails, social media, web surfing, etc.) is stored in phone company, bank, and internet service provider (ISP) records, among other places.[75] The information stored contains not only content (e.g., the text of an email), it also contains transactional (non-content) metadata that describes dates, times, people, and locations.[76] Collectively, this information is referred to as "big data."[77] In this data there is a seemingly infinite amount of information about an individual's

69. Link prediction algorithms actually consider more nuanced topographical features—for example, "Jaccard's coefficient," which reflects link density around a group of individuals. Fire et al., *supra* note 56, at 10:8–10:9.

70. *See* Zheleva et al., *supra* note 57, at 103–04 (considering pet breed as a feature).

71. However, link prediction algorithms are not limited to the most glaring of features, and they will often consider hundreds of descriptive features. *See, e.g.*, Ben Taskar et al., *Link Prediction in Relational Data*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 16, at 659, 663 (2003) (utilizing almost 1,000 features).

72. *See* Indika Kahanda & Jennifer Neville, *Using Transactional Information to Predict Link Strength in Online Social Networks*, *in* PROCEEDINGS OF THE 3RD INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA 74–81 (2009) (achieving 87% AUC).

73. Fire et al., *supra* note 56, at 23 (stating that even when people are "two hops from each other," predictions are able to distinguish between friends and nonfriends).

74. Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 354 (2015).

75. *Id.* at 355–56.

76. *Id.*

77. *Id.* at 352.

social network topography, as well as descriptive information. However, while all the information may be available, taking it as it comes would be akin to drinking from a fire hose. Link prediction system designers must perform a balancing act between capturing the most data possible and doing so in a way that will be both accurate and usable. As discussed below, the SEC has access to a tremendous wealth of transactional metadata. And therein lies the difficulty.

Fortunately, link prediction researchers are on the job. Data fusion is the process of synthesizing data from multiple sources so that the information is not only digestible, but also achieves synergistic effect.[78] This is done by compiling disparate data into matrices and assigning weight to data types and combinations in order to create a standardized, consolidated output.[79] This process not only provides a means of multi-source information management, but can also be used to render more refined factors that are tailored to the link prediction task.

### *D. Link-Strength Prediction*

As rewarding as building a better link predictor might be, those who study link prediction recognize that finding a link does not alone guarantee a practical result.[80] This fact has led some legal-policy-focused writers to doubt the successful application of link prediction in criminal investigation contexts. "People are related to criminals without being criminals themselves," one author writes.[81] "In a wired world, people are more closely connected than we think."[82] This poses, he argues, severe risk of including false-positive links at only a few degrees of separation.[83] However, researchers have not been so easily discouraged by the well-known colloquialism that everyone is six degrees from everyone else. Studies also focus on link-strength prediction, which estimates link importance rather than

78. Kahanda & Neville, *supra* note 72, at 75.

79. For example, imagine one is trying to detect who is the better friend of X, Y, or Z. Phone records may show two calls from X to Y and one from X to Z, while no emails have been sent between X and Y, but two have been sent between X and Z. Now assume that test data shows that not only are phone calls a stronger indication of friendship than emails, but that the presence of phone calls combined with an absence of emails indicates a very close friendship. If the information is weighted accordingly, we can conclude that X and Y have a higher degree of friendship than X and Z. The comparative degree of friendship, then, can be used as a factor for link prediction in lieu of, or in addition to, the raw data.

80. *See* Kahanda & Neville, *supra* note 72, at 74 (pointing out the issue of "spurious relationships").

81. Ferguson, *supra* note 74, at 409.

82. *Id.* (citing, among other things, a study showing that among Facebook users the "average number of acquaintances separating any two people in the world was not six [as is often said colloquially] but 4.74").

83. *Id.* ("Thus, a link analysis . . . can cast a very wide net, accidentally capturing many people who are only suspicious by this loose, associative relationship.").

existence.[84] Not coincidentally, just as links can be predicted through topographical and descriptive information, so too can that information be used to predict link strength.[85]

But while factors like frequency of contact between two people certainly indicate a strong connection in an objective sense, when the goal is to filter connections among co-conspirators from innocent friendships, a different set of criteria is required. One particular technique has been touted for its usefulness in flagging the sort of links one might expect in criminal networks. It is based on the theory that the features of criminal networks are naturally unlike the features of normal social networks. Thus, connections that would otherwise be irrelevant (for example, a connection to one's grandmother) can be used to infer *a contrario* criminal links.[86]

## IV. Where the Rubber Meets the Road

In an insider trading investigation, link prediction would naturally be used to predict a suspect's inside connections by analyzing information investigators could expect to easily obtain. Thus, given that the technological restraints preventing practical applications of link prediction are fading, one must now ask, how much information can investigators expect to easily obtain? The answer is an astronomical amount.

Today, the easiest way to learn about a stranger is through Google. People regularly leave a trail of personal information online—for example, in public postings on social media, in work profiles on company websites, and in other various and sundry locales across the internet. However, to flesh out a social network to the degree required for link prediction, more data is needed. Company records, phone-company and ISP records, and personal-computer and cell-phone memory contain a treasure trove of data—both transactional and content. Getting access to that information poses surprisingly few legal hurdles. First, under the third-party doctrine exception to the applicability of the Fourth Amendment, transactional data held by phone companies and ISPs can be accessed with little fanfare, often with only an administrative subpoena. Second, obtaining information housed in company records, personal computers, and cell phones requires only a warrant.

But despite an optimistic forecast, courts and the Constitution are somewhat hostile to the use of statistical data in certain contexts. First, there is some debate as to whether statistical evidence alone—such as a statistical probability that a trade was an illegal insider trade—can be the basis for a warrant. In addition, there are potential due process concerns that arise from

84. Kahanda & Neville, *supra* note 72, at 74.

85. *Id.* at 74–75.

86. Berlusconi et al., *supra* note 61, at 18.

being investigated and prosecuted using proprietary probabilistic algorithms. Finally, there is considerable opposition to using probabilistic evidence to support a verdict. In the end, these restrictions define the worst-case scenario for the SEC, should it decide to adopt link prediction technology.

### *A. Acquiring Transactional Data: Administrative Subpoenas, the Constitution, and the Stored Communications Act*

When building the topographical map emanating from a suspected insider trader, transactional data is king. While transactional data is certainly contained on private devices, and the seizure of those devices would require a warrant (discussed in Section C), a subpoena to produce those records does not. Furthermore, a sea of transactional data is kept in third-party storage (e.g., phone company records), vulnerable to administrative subpoena per the terms of the Stored Communications Act (SCA).

*1. The SEC's Power to Issue Administrative Subpoena.*—The SEC is empowered to commence any investigation "it deems necessary to determine whether any person has violated, is violating, or is about to violate any provision of this chapter, the rules or regulations thereunder, the rules of a national securities exchange[, etc.] . . . ."[87] To further its investigations, it is authorized to compel the production of documents through administrative subpoena.[88]

For a court to give a subpoena the weight of law, the Supreme Court has stated that if "the investigation [is] conducted pursuant to a legitimate purpose, . . . the inquiry [is] relevant to the purpose, . . . the information sought is not already within the [agency's] possession, and . . . the administrative steps required by [statute] have been followed," the court should defer to the agency notwithstanding evidence of abuse.[89] Although the Supreme Court last addressed the issue in 1946, lower courts have continued to adhere to this policy of wide deference for administrative subpoenas.[90] For example, the Second Circuit in *In re McVane* held that "[a]n affidavit from a government official is sufficient to establish a prima facie showing that [the]

87. 15 U.S.C. § 78u(a)(1) (2012).

88. § 78u(b) reads,

> For the purpose of any such investigation, or any other proceeding under this chapter, any member of the Commission or any officer designated by it is empowered to administer oaths and affirmations, [subpoena] witnesses, compel their attendance, take evidence, and require the production of any books, papers, correspondence, memoranda, or other records *which the Commission deems relevant or material to the inquiry*.

*Id.* § 78u(b) (emphasis added).

89. United States v. Powell, 379 U.S. 48, 57–58 (1964).

90. *See, e.g.*, *In re* McVane, 44 F.3d 1127, 1135–36 (2d Cir. 1995) (acknowledging the court's limited role in enforcing an administrative subpoena).

requirements [for subpoena] have been met."[91] The Second Circuit continued,

> The courts' role in a proceeding to enforce an administrative subpoena is "extremely limited." We defer to the agency's appraisal of relevancy, which "must be accepted so long as it is not obviously wrong." . . . The relevance of the sought-after information is measured against the general purposes of the agency's investigation, "which necessarily presupposes an inquiry into the permissible range of investigation under the statute." . . . We have interpreted relevance broadly.[92]

Under the Second Circuit's policy of deference, the requirement that the information sought be relevant to an investigation is a particularly low bar. Assuming investigators can accurately pinpoint trades too fortuitous to have occurred without reliance on nonpublic information, it is hard to imagine that traders' communications records would not be relevant to an investigation into a violation of the Securities Exchange Act.

*2. A Lack of Constitutional Constraint.*—Regarding any Fourth Amendment constraint on administrative subpoena power, the Supreme Court stated in 1946 that demands to produce records "present no question of actual search and seizure, but raise only the question whether orders of court for the production of specified records have been validly made."[93] After all, in a demand for production, "[n]o officer or other person [seeks] to enter . . . premises against [a person's] will, to search them, or to seize or examine their books, records or papers without their assent . . . ."[94] Moreover, for those subpoenas that have been taken to court for enforcement, the producing party has "adequate opportunity to present objections."[95] Thus, at most, the Court concluded, "the Fourth, if applicable, . . . guards against abuse only by way of too much indefiniteness or breadth" or by not being "relevant" to a legitimate inquiry.[96]

As for protesting a subpoena issued for data stored on third-party servers (e.g., phone records and emails), the Fourth Amendment offers no protection at all. The third-party doctrine, which states that one gives up any reasonable expectation of privacy when one gives information to a third party, leaves the subject of the transactional records without a constitutional leg to stand on.[97]

---

91. *Id.* at 1136.
92. *Id.* at 1135–36.
93. Okla. Press Pub. Co. v. Walling, 327 U.S. 186, 195 (1946).
94. *Id.*
95. *Id.*
96. *Id.* at 208.
97. SEC v. Jerry T. O'Brien, Inc., 467 U.S. 735, 743 (1984) states,

The originator and recipients of the communications are not even required to have notice of the subpoena.[98]

As for the Fifth Amendment's protections against self-incrimination, the Supreme Court has stated, "in so far as [subpoenas] apply merely to the production of . . . records . . . [it] seems to be that the Fifth Amendment affords no protection by virtue of the self-incrimination provision."[99]

*3. Statutory Constraints on Administrative Subpoenas: the SCA.*—In the name of privacy, Congress enacted the Stored Communications Act, introducing statutory restraints on access to electronically stored third-party data.[100] Accordingly, the SEC must comply with the terms of the SCA if it is to gain access to the wealth of transactional data held by phone companies and ISPs in order to satiate its need for topographical data. Fortunately for the SEC, the SCA provides a few exceptions,[101] including a mandate that government entities be given access to non-content data when presented with an administrative subpoena.[102] Again, giving notice to the subject of the investigation is not required.[103]

*4. Final Thoughts on Transactional Data Acquisition.*—When it comes to suspects' transactional data, the data most valuable to the success of link prediction systems, the SCA opens the door to a veritable smorgasbord. Armed with only a red flag marking a suspicious trade and a stack of

---

> It is established that, when a person communicates information to a third party even on the understanding that the communication is confidential, he cannot object if the third party conveys that information or records thereof to law enforcement authorities. . . . These rulings disable respondents from arguing that notice of subpoenas issued to third parties is necessary to allow a target to prevent an unconstitutional search or seizure of his papers.

98. *Id.* at 742–43.

99. *Walling*, 327 U.S. at 208. However, the viability of a Fifth Amendment *due process* claim is discussed in a later section.

100. 18 U.S.C. § 2702 provides in part,

> a person or entity providing an electronic communication service to the public shall not knowingly divulge to any person or entity the contents of a communication while in electronic storage by that service; and . . . a provider of remote computing service or electronic communication service to the public shall not knowingly divulge *a record or other information pertaining to a subscriber to or customer* of such service . . . *to any governmental entity*.

18 U.S.C. § 2702 (2016) (emphasis added).

101. *Id.* § 2702(b)–(c).

102. *Id.* § 2703(c)(2) provides that companies storing communication data,

> shall disclose to a governmental entity . . . when the governmental entity uses an administrative subpoena . . . [a subscriber's] name; address; local and long distance telephone connection records, or records of session times and durations; length of service (including start date) and types of service utilized; telephone or instrument number or other subscriber number or identity . . . .

103. *Id.*

subpoenas, investigators should be able to build a substantial topographical map surrounding the suspected insider trader without the trader's knowledge. When that network is coupled with a list of company employees and business relationships, a well-trained link prediction system should be able to detect simple, and potentially complex, hidden links, telling investigators where to start digging.

### *B. Fourth Amendment Seizures, and Probabilistic Probable Cause*

Unlike for transactional data, the SCA provides increased protection for stored content, such as the body of an email. But even then, protections are minimal. The statute requires a warrant to access content in storage for less than 180 days,[104] but content stored for longer may be requested through administrative subpoena with prior notice.[105] This sweeps older emails, voicemails, text messages, and social media communication into the same easy-access category as transactional data. Nevertheless, during an investigation, waiting 180 days may not be feasible. Moreover, other important information—for example, electronic address books—may only be stored locally on privately owned devices, requiring a warrant to access. With warrants come more robust protections from the Fourth Amendment. This in turn raises concerns about finding probable cause based solely on probabilistic statistical data, such as a statistical probability that a trade was illegal.

*1. The Requirements of Probable Cause.*—The crux of reasonable search and seizure analysis under the Fourth Amendment is probable cause.[106] The probable cause standard exists to accommodate both "individual liberty and public security/crime prevention."[107] The equivalent standard for law enforcement stops is "reasonable suspicion."[108] While the threshold required for reasonable suspicion is generally agreed to be less than probable cause, the considerations are the same. As the Constitution protects against unreasonable search and seizure, the inquiry turns on objective reasonableness.[109] This in turn requires a look at the "totality of the circumstances" of the particular case to see whether there is a prospective

104. *Id.* § 2703(a).

105. *Id.* § 2703(b).

106. U.S. CONST. amend. IV ("[N]o Warrants shall issue, but upon probable cause."); *Katz v. United States*, 389 U.S. 347, 357 (1967) ("[S]earches conducted outside the judicial process, without prior approval by judge or magistrate, are per se unreasonable under the Fourth Amendment—subject only to a few specifically established and well-delineated exceptions . . . .").

107. Erica Goldberg, *Getting Beyond Intuition in the Probable Cause Inquiry*, 17 LEWIS & CLARK L. REV. 789, 795 (2013).

108. The phrase "reasonable suspicion," however, was first articulated in a dissent. Terry v. Ohio, 392 U.S. 1, 37 (1968) (Douglas, J., dissenting).

109. Ferguson, *supra* note 74, at 341–42.

"particularized and objective basis" for a search or seizure.[110] A search that was not justified *ex ante* cannot be legitimated by what was found during the search.[111]

Although the Court describes a seemingly unrestricted totality of the circumstances test to satisfy the Fourth Amendment's reasonableness requirement, one particular type of evidence must be present: there must be some evidence of wrongdoing that is particular to the individual.[112] For example, in *Florida v. J.L.*,[113] an anonymous call "reported to the Miami-Dade Police that a young black male standing at a particular bus stop and wearing a plaid shirt was carrying a gun."[114] Based only on this tip, police searched a young black man who was wearing a plaid shirt.[115] Finding a gun, they arrested him.[116] The Supreme Court held that the anonymous tip, which contained no information particular to that individual—only that a black man in a plaid shirt had a gun, was insufficient for reasonable suspicion.[117] Conversely, in *Alabama v. White*,[118] the Supreme Court held that a tip that included the suspect's name, vehicle, and destination was sufficient to satisfy the Fourth Amendment.[119] The individualized-evidence requirement creates problems for the use of statistical information to clear Fourth Amendment constraints. This is discussed below.

*2. Using Probabilities for Probable Cause*.—There are grave concerns about the ability of probabilistic evidence to support a verdict (discussed below). However, the core inquiry behind the Fourth Amendment's protections against unreasonable search and seizure dovetails nicely with

110. United States v. Arvizu, 534 U.S. 266, 273 (2002).

111. Ferguson, *supra* note 74, at 339.

112. Interestingly, this is not true for stops made pursuant to a policy to stop all passersby—for example, a DWI checkpoint on a public road. In *Michigan Department of State Police v. Sitz*, the Supreme Court, without acknowledging an individualized-evidence requirement, held DWI checkpoints reasonable under the Fourth Amendment. 496 U.S. 444, 455 (1990). In doing so, it performed a balancing test, weighing the strong state interest in preventing drunk driving, the proceduralized nature of the checkpoint, and evidence of drunk drivers on Michigan roads against the minimally invasive nature of the stop. *Id.* Justice Brennan, however, took issue with the majority divorcing checkpoint stops from traditional *Terry*-stop requirements, stating that "individualized suspicion is a core component of the protection the Fourth Amendment provides against arbitrary government action." *Id.* at 457 (Brennan, J., dissenting). Here, however, we discuss individual investigations—not checkpoints. Accordingly, *Sitz* should be noted as a strange, but ultimately irrelevant, exception to the individualized-evidence requirement.

113. 529 U.S. 266 (2000).

114. *Id.* at 268.

115. *Id.*

116. *Id.* at 268–69.

117. *Id.* at 271 (stating that there was no evidence specific to the defendant).

118. 496 U.S. 325 (1990).

119. *Id.* at 327, 331–32.

probabilistic information.[120] Under the probable cause or reasonable suspicion standard, the inquiry focuses on the likelihood of the existence of a fact,[121] a purely probabilistic notion. Nevertheless, many have argued the merits and strikes against using solely probabilistic evidence to satisfy the Fourth Amendment.[122] Moreover, scholars have questioned whether big-data predictions can satisfy the individualized-evidence requirement.[123] To date, there is no word from the Supreme Court as to whether probabilistic evidence alone can satisfy the Fourth Amendment.

*a. The Argument for Probabilistic Probable Cause.*—It is hard to imagine a decision regarding a future outcome that is not made based upon a probability. Recent research on decision-making suggests that humans come to conclusions by entertaining multiple narratives, reevaluating their likelihood based on information as it is received.[124] Even older theories of human decision-making were based around the idea that humans make decisions based on probabilities that are continuously recalculated using new information.[125] Some have argued that people's resistance to relying on purely probabilistic information to make legal determination stems not from the inferiority of probabilistic information; it stems from the fact that the chance of being incorrect is plainly apparent—not buried beneath layers of human intuition and judgment.[126] To that end, an argument for using probabilistic evidence is that one must confront error rates as they come, and if they are unsatisfactory, pursue a more accurate result.

Moreover, we routinely use probabilistic evidence to satisfy the Fourth Amendment without regarding it as such. For example, an arrest may be premised on a drug dog's sniff or a breathalyzer reading even though dogs identify false-positive smells and breathalyzers have margins of error.[127]

120. Max Minzner, *Putting Probability Back into Probable Cause*, 87 TEXAS L. REV. 913, 957–58 (2009) (discussing probability as being integral to probable cause).

121. Brinegar v. United States, 338 U.S. 160, 175–76 (1949) ("In dealing with probable cause . . . as the very name implies, we deal with probabilities.").

122. Andrew Guthrie Ferguson, *Predictive Policing and Reasonable Suspicion*, 62 EMORY L. J. 259, 298–99 (2012).

123. Ferguson, *supra* note 74, at 330.

124. Minzner, *supra* note 120, at 953–54.

125. Dan Simon, *A Third View of the Black Box: Cognitive Coherence in Legal Decision Making*, 71 U. CHI. L. REV. 511, 538–39 (2004).

126. Daniel Shaviro, *Statistical-Probability Evidence and the Appearance of Justice*, 103 HARV. L. REV. 530, 533–34 (1989).

127. *See, e.g.*, Illinois v. Caballes, 543 U.S. 405, 410 (2005) (Souter, J., dissenting) (presuming that a drug dog's sniff can be the sole basis for a search); United States v. Waltzer, 682 F.2d 370, 372 (2d Cir. 1982) (holding that a drug dog's alert "itself establish[es] probable cause, enough for [an] arrest"); *see also* Goldberg, *supra* note 107, at 791 ("Whether or not probable cause exists to issue the warrant depends largely on [the dog's] reliability, which can be quantified based on [the dog's] error rate in detecting drugs."). Given holdings like *Caballes* and *Waltzer*, it is hard to believe

Consequently, it is illogical to declare certain probabilities an affront to the Constitution while others remain widely used in practice.

Finally, the use of quantitative empirical data reduces reliance on intuition, which infuses the constitutionally mandated question of reasonable likelihood with subjective valuations. Even the standard of probable cause itself suffers from subjectivity issues. For example, in one study, a professor asked federal judges to quantify probable cause.[128] The results ranged from ten to ninety percent.[129] Adding objectively determined probabilities to the mix would help ameliorate what is, at its core, an error-prone inquiry.

*b. The Argument Against Probabilistic Probable Cause*.—The main strike against probabilistic probable cause is that the Supreme Court has emphatically stated that "[t]he probable-cause standard is incapable of precise definition or quantification into percentages"[130] because, at its core, "it deals with probabilities and depends on the totality of the circumstances."[131] In other words, it is to be based upon "the factual and practical considerations of everyday life on which reasonable and prudent men, not legal technicians, act."[132] Again, in other words, the probability component is human-derived—meant to be filtered through the reasonable actor.[133] Only at that point is the inquiry reduced to probability alone and, even then, only of a sort that is unrestrained by a need to assign a particular percentage. For that reason, one might argue a probable cause inquiry based solely on probabilistic information undermines the role of the reasonable actor by reducing what was intended to be an unconstrained evaluation of the totality of the circumstances into a pure numbers game.

Nevertheless, the Supreme Court's refusal to constrain probable cause to a number does not necessarily mean that probabilistic information cannot be considered nor that it cannot be the only thing considered when it is all that is available. Statistical evidence, rather than being taken as gospel, should be evaluated by investigators by questioning the credibility of its source and its relative value to the investigation, among other things.

---

courts would question arrests made on the basis of positive breath tests on the grounds that the test results are probabilistic.

128. Ric Simmons, *Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal Justice System*, 2016 MICH. ST. L. REV. 947, 987–88 (2016).

129. *Id.*

130. Maryland v. Pringle, 540 U.S. 366, 371 (2003).

131. *Id.*

132. Brinegar v. United States, 338 U.S. 160, 175 (1949).

133. *See* Ferguson, *supra* note 74, at 405 (noting the Supreme Court's refusal to "quantify" the "standard of reasonable suspicion" because police officers, operating as "reasonable and prudent" persons, assess the standard).

*c. The Individualized-Evidence Requirement in this Context.*—A frequent complaint about probabilistic probable cause is that it can be used as a means to lawfully conduct searches and seizures based purely upon innocent behavior correlated with illegal conduct under circumstances outside the control of the subject, creating scenarios in which one's privacy rights may be deprived simply for engaging in innocent conduct.[134] It could be argued that seeking a warrant based upon an unusual trade as determined by statistical analysis of market movements constitutes such an ill. Indeed, scholars have argued that big-data predictions are an insufficient independent basis to satisfy the Fourth Amendment.[135] This, however, is an oversimplification.

While the idea of using a sweeping generality—like the street corner example above—as carte blanche to search anyone who falls within its umbrella seems patently unreasonable, in other contexts, we are perfectly comfortable relying on correlations built upon information gathered from sources other than the suspect to infer illegal conduct. As mentioned above, there are few qualms about using a drug-sniffing dog as the basis for a lawful search. However, a dog is not trained by smelling a particular suspect with and without drugs. The dog alerts its officer to a smell that has, in the past, successfully been associated with drugs. In essence, the suspect is searched for smelling like other circumstances the dog has been exposed to. The same can be said for breathalyzers. Because they are not calibrated by measuring the alcohol content of a particular suspect's breath, they can only speak to the similarities between the suspect and the device's other subjects.

*d. Final Thoughts on Probabilistic Probable Cause*.—Given that we in fact do use probabilistic evidence to clear Fourth Amendment burdens, the rule cannot be that probabilistic evidence alone is insufficient. Thus, the permissible use of probabilistic evidence must turn on other factors. One factor separating the street corner example from a breath test is the degree of accuracy. Although 90% accuracy seems high in most circumstances, perhaps it is insufficient when invading one's privacy based purely on ostensibly innocent conduct. Another separating factor is how central the conduct being scrutinized is to the offense in question.[136] The alcohol on one's breath speaks directly to intoxication whereas one's presence on a street corner by itself does not implicate drug possession. Yet another, some

134. For example, police officers searching every person passing by a particular street corner at night based on evidence that 90% of all people on that particular street corner at night are carrying drugs.

135. Simmons, *supra* note 128, at 950 (stating in no uncertain terms that big data is unable to provide the type of individualized information required to satisfy the Fourth Amendment).

136. *See* Ferguson, *supra* note 74, at 338–39 (noting that "[s]uspicion must be particularized," relating to the "current criminal activity").

have argued, is the amount of individualized information used as the basis for the prediction.[137] However, this does not explain the acceptance of breathalyzers, which only take a single sample.

If these differences are meaningful, there should be no reason why statistical evidence of a trade too fortuitous to have been legal cannot pass muster under the Fourth Amendment, allowing investigators to obtain a warrant to search the personal devices of suspected insider traders. Provided that market-monitoring systems are sufficiently accurate, the conduct being scrutinized is the trade suspected of yielding fraudulent gains, and it is being critiqued in part by the magnitude of those gains.[138]

### *C. Due Process Concerns*

Generally, the Fourth Amendment is said to displace Fifth Amendment due process claims arising from searches or seizures because the Fifth Amendment is not intended to serve as a vortex for substantive rights housed elsewhere in the Constitution. But some have argued that the Fourth Amendment might not always dictate the constitutional propriety of pulling data.[139] There are some circumstances where employing algorithmic data in the investigation and prosecution of a crime would implicate conceptions of fairness rather than reasonableness. In those circumstances, where the right being protected is in fact the right to due process, investigators' efforts may be constrained, not by the Fourth Amendment, but by the Fifth.

The foremost due process concern is a lack of transparency, which may hide a multitude of sins. As one scholar decried, with an algorithm "[t]here is no notice, no opportunity to be heard, no confrontation with adverse evidence, and no reason given—only a result."[140] So too has Justice Ginsburg expressed concern, writing,

> Inaccuracies in expansive, interconnected collections of electronic information raise grave concerns for individual liberty. The offense to the dignity of the citizen who is arrested, handcuffed, and searched on a public street simply because some bureaucrat has failed to maintain

137. *Id.* at 330.

138. Of course, this is not the only level of abstraction from which to view the inquiry. One might just as easily recast the inquiry as scrutinizing profit—a perfectly legal and desirable outcome. However, this does not serve to distinguish it from the breathalyzer example. Having alcohol on one's breath is not itself illegal. Illegality comes from the level of intoxication or the amount of alcohol in one's blood. Breathalyzers attempt to predict blood-alcohol content by using, as their sole criterion, the alcohol on a subject's breath. Similarly, market analysis uses profit as one of many factors to predict illegal trading behavior.

139. Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process*, 40 GA. L. REV. 1, 45–46 (2005).

140. *Id.* at 45.

> an accurate computer data base is evocative of the use of general warrants that so outraged the authors of our Bill of Rights.[141]

Indeed, the questions of fairness are many: If it took a computer two hours to do what a team of investigators would have taken weeks to do, how can defendants hope to unwind the steps taken by an algorithm to mount a defense against the information it found valuable? Do defendants receive due process when the building blocks of their criminal cases are only rebuttable through a battle of experts? Moreover, police are not allowed to recklessly withhold exculpatory evidence from an affidavit for a warrant.[142] What if a program was recklessly designed to miss certain exculpatory evidence? How would anyone know that such evidence was not taken into account?

There are no answers to these questions.[143] However, each implicates the fundamental sense of fairness that undergirds so much of our criminal process. For that reason, the Fifth Amendment might provide a means to protest the use of such evidence despite the general rule that searches and seizures are best analyzed under the Fourth. As to the extent defendants should expect to find shelter in the Fifth Amendment, at this point, it is too early to tell.

### *D. Supporting a Verdict on Analytics Alone?*

Although there has been no word from the United States Supreme Court, various courts have held that probabilistic evidence alone cannot support a verdict.[144] This somewhat-accepted rule is derived from the proposition that fact finders must determine the existence of facts, not conclude that facts have a probability of existing.[145] To some this may seem absurd—after all, the preponderance of the evidence standard is regularly described as a finding that something was "more likely than not." However, proponents of the rule argue that burden of proof standards exist only to explain the level of

---

141. Herring v. United States, 555 U.S. 135, 155–56 (2009) (Ginsburg, J., dissenting) (internal quotation marks omitted).

142. *See, e.g.*, Whitlock v. Brown, 596 F.3d 406, 410–11 (7th Cir. 2010) (examining whether a policeman recklessly withheld evidence from a particular warrant application).

143. Although, for a very thorough *Mathews v. Eldridge* analysis of big data analytics, see Steinbock, *supra* note 139, at 54–62.

144. For a discussion, see Minzner, *supra* note 120, at 956–57.

145. As an illustration, take for example the seminal case *Smith v. Rapid Transit Inc.*, 58 N.E.2d 754 (Mass. 1945). In that case, the plaintiff was run off the road by a bus. *Id.* at 754–55. Although the plaintiff could not identify the bus company by memory, she proffered a mathematical probability that given the scheduled bus routes, the bus was owned by the defendant. *Id.* Dismissing the case, the court stated, "A proposition is proved by a preponderance of the evidence if it is made to appear more likely or probable in the sense that actual belief in its truth, derived from the evidence, exists in the mind or minds of the tribunal notwithstanding any doubts that may still linger there." *Id.* (internal quotation marks omitted). For a pragmatic argument against this rule, see Shaviro, *supra* note 1266, at 539–40 (arguing, among other things, that requiring jurors to form beliefs introduces inaccuracy to the point that we might be better off with pure probabilities).

confidence one must have in one's conclusion—whether a fact has or has not been proven—not to create a probability threshold to stand in the place of an actual determination.[146] Admittedly, this sounds like splitting hairs. However, this debate makes it difficult to hope that insider trading prosecutors will be able to reliably secure a verdict by proffering only a prediction that a trade was based on insider information and a prediction that the information was obtained from a particular source to satisfy two of the necessary elements of the action.

## Conclusion: The Worst-Case Scenario

Should the SEC adopt link prediction technology as a means to investigate suspicious trades that do not readily appear to have inside connections, the worst-case scenario is that its application will be limited to reconnoitering suspects' networks using only transactional third-party data and the data investigators can procure from willing companies. Investigators will be unable to use market analysis or link predictions as a means to secure a warrant, and prosecutors will be unable to use that data to sustain a verdict. But even with these limitations, the SEC is poised to benefit greatly from link prediction. When faced with a suspicious trade with no discernable connection to an insider, investigators will have a means to conduct a low-cost preliminary case audit before deciding to proceed further. Furthermore, if a link is predicted, investigators will have a direction in which to move. This will allow the SEC to pursue cases that might otherwise have been cut in an *ex ante* triage decision and save investigative resources that can be redirected towards pursuing additional cases. Accordingly, even if the legal bars discussed above prevent extensive reliance on link prediction as a means of enforcement, the SEC is nevertheless poised to take yet another step towards its goal of unrelenting omnipresence.

*Andrew P. Van Osselaer*

146. *See, e.g.*, *Smith*, 58 N.E.2d at 755 (explaining the preponderance of the evidence standard as referring to the jurors' actual belief in the truth of the evidence).