M&A and Innovation: A New Classification of Patents

By Zhaoqi Cheng, Ginger Zhe Jin, Mario Leccese, Dokyun Lee and Liad Wagman*

Recent antitrust debates focus on mergers and acquisitions (M&A), particularly those in which large, established firms acquire startup ventures in nascent technological areas. On the one hand, such M&A may enable a startup to access more resources and strengthen both sides of the deal in innovation, product, and market reach; these synergies can be socially beneficial if they foster innovation and competition. On the other hand, it is of concern that such acquisitions may reduce the profitability of other startups in a similar business or technology area, allow the acquirer to strategically deter or impede the entry and/or growth of competitors, and stifle future innovation. These anti-competitive risks may take place in the form of killer acquisitions (Cunningham, Ederer and Ma, 2021), kill zones (Kamepalli, Rajan and Zingales, 2020), reverse killer acquisitions (Caffarra, Crawford and Valletti, 2020), and entry-for-buyout (Bryan and Hovenkamp, 2020).

While the debates are ongoing, some antitrust authorities have posited towards more enforcement, citing concerns of future competition and innovation in nascent markets.[1] These changes were met with criticisms that the agencies are speculative about what incumbents may have done without the M&As, and that the "I know it when I see it" approach in the US Federal Trade Commission's (FTC) 2022 policy statement on unfair competition may deter healthy competition and disruptive innovations.[2,3]

To resolve these intellectual and policy debates, it is necessary to understand how technological innovations evolve vis-à-vis business dynamics within and across firms, including M&A. To our best knowledge, the related empirical literature tends to emphasize either technological innovations or business dynamics, but rarely investigates how the two intertwine and evolve synchronously.[4]

Drawing a closer tie between technological innovations and business dynamics is essential in addressing whether incumbent acquisitions of startups foster, displace, or deter innovations. This paper develops a methodology that explicitly incorporates the link-

* Cheng: Boston University, zqc@bu.edu. Jin: University of Maryland & NBER, ginger@umd.edu. Leccese: University of Maryland, leccese@umd.edu. Lee: Boston University, dokyun@bu.edu. Wagman: Illinois Institute of Technology, lwagman@stuart.iit.edu. We thank the Washington Center for Equitable Growth and our home universities for financial support. Hyo Kang and attendees at the 2023 AEA meetings provided constructive comments. All errors are ours.

[1] For example, in October 2022, UK's Competition and Market Authority (CMA) ordered Meta to divest of Giphy, the largest supplier of animated GIFs to social networks, which Facebook/Meta acquired two years ago for $315 million (`https://www.gov.uk/government/news/cma-orders-meta-to-sell-giphy`.). In December 2022, the US FTC sued Microsoft to block its acquisition of Activision Blizzard, alleging that the deal may enable Microsoft to withhold content from its gaming rivals and harm competition in the nascent market of cloud gaming (`https://news.bloomberglaw.com/mergers-and-acquisitions/microsofts-fight-for-activision-looks-to-gamings-cloud-future`.

[2] See `https://www.uschamber.com/finance/antitrust/the-ftcs-section-5-policy-statement-effectively-declares-competition-illegal`.

[3] As antitrust agencies are often resource constrained, Jin, Sokol and Wagman (2022) note that the informational infrastructure of the US antitrust system has massively lagged behind the digital economy, which hampers their ability to supervise the technologically-charged business environment. A recent survey finds that practitioners currently perceive the FTC and the Department of Justice (DOJ) as less transparent, less fair, and more combative than previously (Sokol et al., 2022).

[4] In particular, the patent literature associates patent counts, patent quality (e.g., originality, generality, unconventionality), and economic value of patents with innovator attributes, innovation waves, and firm-level statistics. Much of the literature considers innovation within the closed-loop system of US Patent and Trademark Office (PTO) documentation, but does not explore how patents within and across firms relate in terms of innovation directions and business synergy. A separate literature studies M&A and venture investment deals, often treating each firm as a 'black box,' and rarely exploring how firms overlap with each other's patent portfolios, nor how M&A may affect the future development of these portfolios.

ages between US patent filings and business ownership. Leveraging the US PTO's Cooperative Patent Classifications (CPC) (which encode technological relationships among patents) and patent assignee id (which may capture a clustering of business strategies within each firm), we add a new layer of 'Tech-Business Zone' (TBZ) classification on top of the CPC system. The intuition is simple: if *most* firms that file patents classified under CPC A also file patents classified under CPC B, there are reasons to believe that A and B are related in business, even if they are distinct in technological content. This relationship, if sufficiently strong, suggests that A and B may belong to the same TBZ due to business affinity, even if we do not know how these affinities translate into product, supplier relationship, and other business functions of the patent owner.

Section I describes how we use historical US patent data (2000–2009) to define 172 TBZs and link each patent and patent assignee to the zones on a probabilistic basis. This allows us to describe a firm's patent portfolio at any time as a vector of patents per zone. To showcase the potential utility of combining patent and business ownership data, Section II merges the 2010-2019 US patent data with 2010-2021 M&A data from Standard & Poor's (S&P). We then summarize the connections among acquirers and targets based on the zones defined in Section I. Section III concludes with a research agenda that our zone classification can enable to study.

I. Zone extraction via patents alone

US PTO's PatentsView platform (henceforth PatentsView) records all US patents granted from 1976 to the present, including patent ID, assignee, issue date, CPC symbols, and forward and backward citations. We merge PatentsView with the Google Public Patent database to acquire auxiliary patents' priority date, which refers to the earliest date when a patent application is filed in respect of the focal invention.

The key intuition of our TBZ classification is that patent data contains important connections among CPCs in the patents filed by the *same* patent assignee. To capture these connections, we define an assignee's patent filing portfolio as a collection of CPCs under which the focal assignee filed patents during a specific period. Treating each portfolio as a 'document' and each CPC as a 'word' in the document, we can use machine-learning topic models such as Latent Dirichlet Allocation (LDA) to uncover 'topics' as TBZs.

In particular, LDA is a hierarchical Bayesian model widely used for the discovery of hidden semantic structures in text corpora. Under a document–topic–token hierarchy, a document (e.g., a news article) covers a finite number of *topics* with different proportions (e.g., an article may be 80% about *politics* and 20% about *healthcare*), and each topic is described by a few *words* semantically coherent to each other (e.g., *congress* and *bill* for the topic of politics, *diabetes* and *insulin* for healthcare). A topic model aims to find a set of interpretable topics that best constitute the documents in a training corpus, while ensuring the collections of words that describe each topic to be semantically coherent to each other. In our context, each topic (TBZ) is a distinctive cluster of words (CPCs) that appear coherently in the same document (an assignee's patent portfolio). Statistically, the model can identify topic A from topic B in an unsupervised way because the words within each topic tend to co-appear in the same documents while the words in different topics do not co-appear as often.

In theory, we can use all patents filed during a specific period to define TBZs. However, ownership changes due to M&As in a period automatically contribute to the zone definition, which implies that using the zone definition to study how an acquirer and a target compare in zone coverage would generate incorrect inferences. To address this issue, we use the 2000-2009 patent data as training data for the zone definition. Doing so provides a snapshot of the zone landscape *before* we begin observing M&A deals in the S&P data. This way, we apply the pre-fixed zone definition to every firm's patent portfolio between 2010 and 2019, and describe their differences and overlap.

More specifically, to prepare the training

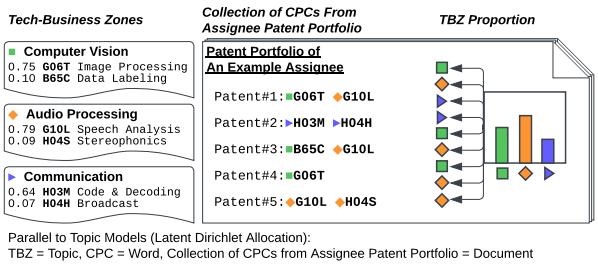

FIGURE 1. STYLIZED ILLUSTRATION OF TECH-BUSINESS ZONE (TBZ) DEFINITION

data we take a subset of PatentsView including only patents with priority dates between 2000 and 2009. We filter out assignees whose patent filings span only one CPC subclass, and the subclasses with less than one filing per year (e.g., E02C, Ship-Lifting Device; G21D, Nuclear Power Plant).[5] These result in 94,014 assignee filing portfolios over 2,050,443 patent records; each assignee portfolio on average contains 90.74 CPCs and spans 5.37 unique CPC subclasses. We then employ LDA to extract TBZs as 'topics' from all assignees' patent portfolios in the training data, as illustrated in Figure 1.[6]

For LDA training, we employ grid search with cross-validation to identify the optimal hyperparameter based on per-topic UMass coherence scores.[7] The best model yields a list of 172 TBZs. We also manually check the zone-describing CPCs to confirm coherence and relativity.[8]

Applying the trained model to patent data during 2010-2019, we infer the TBZ proportions of each assignee's patent portfolio and individual patent. Table 1 summarizes PatentsView throughout 2010–2019 at the patent and zone levels. As PatentsView includes only granted patents, the priority date summary reflects a natural selection towards fewer patents in later years. As more complex patent applications may be subject to longer examinations, the average backward citations per patent are smaller in later years. The number of forward citations declines by priority date as newer patents have less time to receive citations.[9]

Following Hall, Jaffe and Trajtenberg (2001), we define a patent's originality index as one minus the Herfindahl concentration of the focal patent's backward citations over CPCs. During 2010-2019, the average patent originality by CPC fluctuates between 0.5 and 0.55. Applying the same concept to the TBZ proportions of each patent, the average patent originality by zone has a smaller absolute magnitude than that by CPC, and the two originality indices change consistently with each other, as expected. On average, a zone covers more than 1,000 patents and more than 1,000 unique assignees every two years (except for 2018-2019) but each assignee tends to concentrate its patent filings in 1-2 zones

[5] See `https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html` for the CPC scheme.

[6] By estimating the model parameters (i.e., the zone mixture for each portfolio and the distribution over CPCs) via iterative Gibbs sampling, LDA can identify a set of CPCs that primarily constitute and describe each zone, and estimate the proportions of each TBZ in a patent portfolio. See Blei, Ng and Jordan (2003) for additional details on LDA.

[7] See Mimno et al. (2011) for evaluation of semantic coherence in topic models.

[8] For example, we find that both C10G (Hydrocarbon Oils Cracking) and C10M (Lubricating Compositions) belong to the same TBZ, as they involve products from the same production process. The CPC C11D (Detergent Compositions) and A46B (Brushes) belong to the same TBZ, as both concern household products, despite being considered distinct technical domains. Interestingly, we also find G06T (Image Data Processing) appears in a motor-vehicle related TBZ that contains F16D (Clutches & Brakes) and F16H (Gearing), likely because these are components required for driver-assistance systems. These examples suggest that the data-driven TBZ classification not only clusters CPCs close in the technology space, but also clusters CPCs that are related for underlying business reasons.

[9] As in Hall, Jaffe and Trajtenberg (2005), we are reluctant to impute the observed forward citations into life-long citations because the younger cohorts may have too short a patent life to do so reliably.

Table 1—Summary statistics for entire patent data

By priority date of granted patents	2010-2011	2012-2013	2014-2015	2016-2017	2018-2019	TOT
Patent-level						
# of assignees	51,982	57,934	59,157	49,210	22,976	152,818
# of patents	476,977	544,174	535,519	420,131	165,769	2,142,570
Average # of patents granted per assignee	9.18	9.39	9.05	8.54	7.21	14.02
Average # of backward patent citations	17.27	15.73	14.16	13.60	9.82	14.80
Average # of backward scientific citations	10.34	10.21	8.16	6.37	4.69	8.55
Average # of forward citations	4.30	2.72	1.52	0.66	0.22	2.17
Average patent originality by CPC code	0.547	0.536	0.521	0.512	0.501	0.527
Average patent originality by zone definition	0.398	0.395	0.394	0.393	0.387	0.394
Zone-level						
Median weighted # of patents filed per zone	1561.17	1857.43	1827.92	1345.38	432.97	19999.41
Average # of unique assignees that filed patents per zone	1065.84	1247.13	1316.82	1121.45	494.03	7276.78
Concentration of patent filing by assignees	6491.43	6337.76	6207.59	6078.40	6193.83	6544.32

within every two years.

II. Combining patent and M&A data

We next merge the patent data with 451 Research data—a tech M&A database from S&P Global Market Intelligence—(henceforth, S&P). In the S&P data, each observation is an M&A transaction associated with a change in majority ownership. In total, it covers 46,216 M&A transactions involving 21,039 acquirers recorded between 2010 and 2021. All targets are firms belonging to the Information, Communication and Energy Technology (ICET) space but acquirers can operate in any sector.[10]

We first associate each assignee with its legal entity (e.g., NewsCorp is part of Dow Jones, AWS is part of Amazon) using parent company and/or company alias from S&P. To avoid interpreting the acquisition of a subsidiary of a larger entity as that of the entity itself, we attach the same identifier to any parent and its subsidiaries appearing as an acquirer in S&P, but treat each S&P target as a separate entity. We code any patent filed by a target after the M&A transaction as the acquirer's patent.

The name match is conducted on the universe of the S&P data (2010-2021) and PatentsView (1976–present). Out of the 46,216 M&A deals in the S&P data, we are able to match 37.42% of the acquirers, 16.42% of the targets, and 8.96% of both the acquirer and the target in the same deal. The low match rate is unsurprising, as not all firms have patents even if they operate in the broadly defined technology space.

We further restrict the sample up to 2019 because very few patents filed in 2020-2022 appear in PatentsView due to patent examination times. To help compare the patent stock of acquirers and targets, we drop all entities that have zero patent stock in all years between 2010 and 2019, where patent stock in year t is defined as the total number of granted patents that the entity had filed between $t-20$ and $t-1$. This drops 5.1% of the entities.[11] We also drop few cases where the acquirer and the target carry the same entity name. The above leads to a sample of 2,955 S&P M&A deals during 2010-2019 where both acquirer and target names are matched between S&P and PatentsView and both sides have a non-zero patent stock in at least one year after 2010; in 54% of them, both parties filed patents in 2010-2019.[12]

As one may expect, at the time of M&A, targets are, on average, younger than acquirers, are less likely to be publicly traded, have a smaller patent stock per zone, and hold a more concentrated portfolio in terms of zone coverage. Interestingly, the targets that are never acquirers tend to hold patents that,

[10] Jin, Leccese and Wagman (2022*b*) show that while the M&A data by Refinitiv's SDC covers every sector of the economy, it is less comprehensive than the S&P database for majority control deals involving ICET targets. Jin, Leccese and Wagman (2022*a*) provides additional details on the S&P database.

[11] An entity that has zero patent stock throughout 2010-2019 may appear in the universe of the PatentsView data if it had filed a patent before 1990.

[12] Additional details on merging patent and M&A data are provided in the online appendix.

on average, have a higher number of forward citations and a higher average patent originality by CPC, as compared to firms that acquired. This confirms the prior that targets are smaller, younger, and more innovative in their specialty.[13] Since a thorough examination of a technology-driven M&A deal requires a look into how the acquirer and the target correlate in innovation, we examine how the merging parties overlap in TBZs at the time of M&A. We find that targets are not only active in a subset of acquirer-active TBZs but also span some TBZs where the acquirer is inactive.[14]

III. Conclusion and research agenda

The TBZ classification can be used in many lines of research. For merger reviews, it offers an automated method of identifying relevant parties in the patent tech-business space, which can improve the efficiency of the first step of merger review, and guide subsequent data requests, third-party interviews, and counterfactual analysis.[15] Our methodology enables continuous measures of distance between pairs of zones, including measures quantifying how adjacent a target is to its acquirer in the patent space.

The TBZ structure, by definition, is driven by business ownership information embodied in the raw data of patent filings. Hence, one can directly describe the evolution of technological innovation by examining the evolution of TBZs. A dynamic TBZ classification can help policymakers evaluate historical innovation policies (e.g., a subsidy or the initiation of so-called innovation zones). It can also help (i) identify emerging consolidation trends that cut across markets, even if the markets had historically been linked with distinct zones; (ii) track how incumbents expand to adjacent and farther zones; and (iii) whether such trends have effects on patentable innovations.

For practitioners, TBZs can be helpful in assessing the evolution and trends of patent-business relationships as a business intelligence tool and in potential patent disputes and standard-setting efforts. One can use TBZs to understand how inventors' creative efforts shift within/across zones. Such shifts may be influenced by the patent portfolios of employers, the proximity of other inventors, firms' acquisitions and policy changes, and the laws governing employment migration among firms, such as non-compete laws.

[13] These results are in Table A.1 of the online appendix.

[14] In the online appendix, we offer a detailed discussion of the different measures of overlap we examine (cosine similarity, Jaccard similarity and overlap coefficient).

[15] For example, if the merging parties have significant TBZ overlap, one can further compare them with other firms in the same zone. A closer examination of business affinities with other same-zone firms may help understand whether a similar affinity would materialize in the proposed merger, which business affinity motivates the merger deal, and what effects the affinity may have on future patents and future competition. If the acquirer and the target tend to focus on different TBZs, the classification can help identify other firms in their respective zones—firms which may complement or compete with the merging parties in technological innovation.

REFERENCES

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research*, 3(Jan): 993–1022.

Bryan, Kevin A., and Erik Hovenkamp. 2020. "Antitrust limits on startup acquisitions." *Review of Industrial Organization*, 56(4): 615–636.

Caffarra, Cristina, Gregory Crawford, and Tommaso Valletti. 2020. "How tech rolls': Potential competition and 'reverse' killer acquisitions." *Antitrust Chronicle*, 2(2): 1–9.

Cunningham, Colleen, Florian Ederer, and Song Ma. 2021. "Killer acquisitions." *Journal of Political Economy*, 129(3): 649–702.

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg. 2001. "The NBER patent citation data file: Lessons, insights and methodological tools." *Available at NBER 8498*.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg. 2005. "Market value and patent citations." *RAND Journal of economics*, 16–38.

Jin, Ginger Zhe, D. Daniel Sokol, and Liad Wagman. 2022. "Towards a technological overhaul of American antitrust." *Antitrust*, 37(1).

Jin, Ginger Zhe, Mario Leccese, and Liad Wagman. 2022*a*. "How do top acquirers compare in technology mergers? New evidence from an S&P taxonomy." *International Journal of Industrial Organization*, 102891.

Jin, Ginger Zhe, Mario Leccese, and Liad Wagman. 2022*b*. "M&A and technological expansion." *Available at SSRN 4009215*.

Kamepalli, Sai Krishna, Raghuram Rajan, and Luigi Zingales. 2020. "Kill zone." Available at NBER 27146.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing semantic coherence in topic models." 262–272.

Sokol, D. Daniel, Marissa Ginn, Robert Calzaretta, and Marcello Santana. 2022. "Antitrust mergers and regulatory uncertainty." *Available at SSRN 4295283*.