



ARTICLE

**“I SEE SOMETHING YOU DON’T SEE”:
A COMPUTATIONAL ANALYSIS OF THE
DIGITAL SERVICES ACT AND THE DIGITAL
MARKETS ACT**

Fabiana Di Porto,^{*} Tatjana Grote,^{**} Gabriele Volpi^{***} & Riccardo Invernizzi^{****}

Abstract. In its latest proposals, the Digital Markets Act (DMA) and Digital Services Act (DSA), the European Commission puts forward several new obligations for online intermediaries, especially large online platforms and “gatekeepers.” Both are expected to serve as a blueprint for regulation in the United States, where lawmakers have also been investigating competition on digital platforms and new antitrust laws passed the House Judiciary Committee as of June 11, 2021. This Article investigates whether all stakeholder groups share the same understanding and use of the relevant terms and concepts of the DSA and DMA. Leveraging the power of computational text analysis, we find significant differences in the employment of terms like “gatekeepers,” “self-preferencing,” “collusion,” and others in the position papers of the consultation process that informed the drafting of the two latest Commission proposals. Added to that, sentiment analysis shows that in some cases these differences also come with dissimilar attitudes. While this may not be surprising for new concepts such as gatekeepers or self-preferencing, the same is not true for other terms, like “self-regulatory,” which not only is used differently by stakeholders but is also viewed more favorably by medium and big companies and organizations than by small ones. We conclude by sketching out how different computational text analysis tools, could be combined to provide many helpful insights for both rulemakers and legal scholars.

^{*} Professor of Law and technology at the University of Salento, Italy. This paper is part of a broader interdisciplinary research project on algorithmic disclosure regulation, funded by a three-year PRIN research grant awarded by the Italian Ministry of Research (no. of grant 2017BAPSXF, www.lawandtechnology.it). We are thankful to Prof. Julian Nyarko from Stanford University for providing some of the data we used in the empirical part and to a number of other people for insightful discussions: Dr. Roberto Marseglia, Prof. Daniela Piana, Giuseppe Italiano, Giorgio Monti, the discussants at the Ascola annual conference of 2021. All mistakes remain our own.

^{**} Tatjana Grote, BA in International Relations from the Technical University of Dresden and Member of the Algorithmic Disclosure Regulation PRIN-funded Research project 2017–22.

^{***} Gabriele Volpi, MSc Theoretical physics from the University of Pavia, Manager at DAAT consulting and Member of the Algorithmic Disclosure Regulation PRIN-funded Research project 2017–22.

^{****} Riccardo Invernizzi, BSc in Mathematics (with honors) from the University of Pavia, junior analyst at DAAT consulting and Member of the Algorithmic Disclosure Regulation PRIN-funded Research project 2017–22.

I. Introduction

“It is complex but we are looking forward to it.”¹ This is the closing remark of European Commission Executive Vice-President Margrethe Vestager’s statement when introducing the latest Commission proposals on digital platforms: the Digital Markets Act and the Digital Services Act.²

Without a doubt, managing competition on digital markets is not a simple endeavor. This complexity is reflected by the lively scholarly debate on new rules for digital markets.³ In December 2020, the Commission published what everyone involved in these debates had been looking forward to: twin proposals suggesting many new rules for “online platforms,”⁴ “very large online platforms,” and “gatekeepers.”⁵

These rules might serve as a “blueprint for regulation across the globe,”⁶ for at least two reasons. First, they are not only relevant to EU businesses and consumers, but to businesses around the world, especially well-known US tech companies.⁷

Second, given that US lawmakers, agencies and state attorney generals have been investigating competition on digital platforms recently,⁸ and that new

¹ Margrethe Vestager, *Statement by Executive Vice-President Vestager on the Commission Proposal on New Rules for Digital Platforms*, EUR. COMM’N (Dec. 15, 2020), https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_20_2450.

² Eur. Comm’n, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC*, COM/2020/0825 (Dec. 15, 2020) [hereinafter, *DSA Proposal*; the draft regulation therein, hereinafter *DSA*]; Eur. Comm’n, *Proposal for a Regulation of the European Parliament and the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act)*, COM/2020/0842 (Dec. 15, 2020) [[hereinafter, *DMA Proposal*; the draft regulation therein, hereinafter *DMA*].

³ See, e.g., Pablo Ibáñez Colomo, *Whatever Happened to the ‘More Economics-Based Approach’?*, 11 J. EUR. COMPETITION L. & PRAC. 473, 473 (2020) (discussing the growing shift from the so-called “more economic approach” to ex-ante intervention against big digital platforms in the European legal community). For challenges related to competition law, see generally, ARIEL EZRACHI & MAURICE E. STUCKE, *VIRTUAL COMPETITION: THE PROMISE AND PERILS OF THE ALGORITHM-DRIVEN ECONOMY* (2016); and PHILIP MARSDEN & RUPPRECHT PODSZUN, *RESTORING BALANCE TO DIGITAL COMPETITION—SENSIBLE RULES, EFFECTIVE ENFORCEMENT* 1-87 (2020). On consumer protection and its relation to data protection and competition law, see Wolfgang Kerber, *Digital Markets, Data, and Privacy: Competition Law, Consumer Law, and Data Protection*, 11 J. INTEL. PROP. L. & PRAC. 856 (2016). More recently, UNCTAD, *Competition law, policy and regulation in the digital era*, TD/B/C.I/CLP/57, 28 April 2021, https://unctad.org/system/files/official-document/ciclpd57_en.pdf (accessed 7 July 2021) (providing an excellent and up-to-date overview of the debate).

⁴ There is no perfect overlapping between the legal definitions of “platforms” in the DSA and DMA. In the DSA, the widest concept is that of online “intermediary service,” which covers all services within the scope of art. 1(3), including “online platforms” (providing hosting services), as described in art. 2(1)(h). Conversely, in the DMA, the widest concept is that of “core online platform,” described in art. 2(2), which covers “online intermediation services” (inclusive of application stores), together with other services (for example, cloud computing services, social networking sites, video-sharing platforms, search engines, operating systems, and advertising services).

⁵ See *DSA Proposal*, *supra* note 2; *DMA Proposal*, *supra* note 2.

⁶ Aline Blankertz & Julian Jaurisch, *How the EU Plans to Rewrite the Rules for the Internet*, BROOKINGS. (Oct. 21, 2020), <https://www.brookings.edu/techstream/how-the-eu-plans-to-rewrite-the-rules-for-the-internet>.

⁷ See DSA art. 1(3); DMA art. 1(2).

⁸ See SUBCOMM. ON ANTITRUST, COM. & ADMIN. L., H. COMM. ON THE JUDICIARY, 116TH CONG., *INVESTIGATION OF COMPETITION IN DIGITAL MARKETS* (2020), https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf; see, e.g., *Complaint for Injunctive and Other Equitable Relief, Fed. Trade Comm’n v. Facebook, Inc.*, No. 1:20-cv-03590 (D.D.C. Dec. 9, 2020); *Complaint, New York v. Facebook, Inc.*, No. 1:20-cv-03589 (D.D.C. Dec. 9, 2020); *Complaint, United States v. Google LLC*, No. 1:20-cv-03010-APM (D.D.C. Oct. 20, 2020).

antitrust laws passed the House Judiciary Committee as of June 11, 2021,⁹ the newly proposed EU rules might have some visible repercussions on the US legal landscape. To name just a few: If the acts are adopted, “very large online platforms,”¹⁰ will need to *provide meaningful information* on how they manage, (DSA art. 23), and *rank* their content.¹¹ The DMA addresses platforms designated as “gatekeepers”¹² with a list of dos and don’ts. For instance, it prohibits self-preferencing,¹³ while (obliging them to enable multi-homing and interoperability).¹⁴

As can be seen from these examples, terms and concepts like “gatekeepers,” “self-preferencing,” and “interoperability” play central roles in designing new rules for online platforms. Inherently, this comes with the challenge of defining these new concepts to make them legally operable.¹⁵ At the same time, there might be some confusion on how exactly some more “traditional” qualifiers, e.g., “dominant” position or “anti-competitive” conduct, apply to digital services.¹⁶

Consequently, a large part of the debate on new rules for digital markets prior to the publication of the EC twin proposal has centered around how to define or understand certain essential concepts.¹⁷

⁹ In fact, the US House Judiciary Committee only recently passed six acts dealing with this very topic. Most notably, H.R. 3849, ACCESS Act of June 11, 2021, 117th Congress (2021-2022), <https://www.govtrack.us/congress/bills/117/hr3849> (Augmenting Compatibility and Competition by Enabling Service Switching) focuses on interoperability and data portability; H.R. 3816 (American Choice and Innovation Online Act) of June 11, 2021, 117th Congress (2021-2022), <https://www.govtrack.us/congress/bills/117/hr3816> addresses self-preferencing and other anti-competitive conducts (all accessed 3 August 2021).

¹⁰ Some online platforms are subject to additional obligations: these are the “very large online platforms,” defined as those with at least 45 million monthly active users in the EU. See DSA art. 25.

¹¹ DSA art. 29 requires online platforms to disclose the main parameters used in recommender systems so as to prevent self-preferencing practices in rankings.

¹² In the DMA art. 3, some providers of core online services may be designated as “gatekeepers”. There are two different mechanisms for this designation. The default one is based on three quantitative criteria set out in DMA art. 3(2): (1) a turnover equal to or above EUR 6.5 billion or an average market capitalization of at least EUR 65 billion; (2) more than 45 million monthly active end users, and more than 10,000 yearly active business users established in the Union; (3) the second criteria being met for three consecutive financial years. If a platform fulfills all three criteria, it must notify the Commission (DMA art. 3(3)) and will be designated as a gatekeeper within 60 days unless it presents proof that it does not match the description of a gatekeeper in DMA art. 3(1) despite fulfilling the criteria in DMA art. 3(2), DMA art. 3(4). Thus, that a company is a gatekeeper is based on a rebuttable presumption. There is a second mechanism to establish if a company is a gatekeeper. If the thresholds are not met, the provider of a “core platform service” can still be identified as a gatekeeper, DMA art. 3(6), after a market investigation by the Commission, DMA art. 15. Hence, not every very large platform is a gatekeeper, but it is likely that every gatekeeper will also be a very large online platform. See DMA art. 3(2)(b).

¹³ To be precise, the DMA obliges gatekeepers to “refrain from treating more favorably in ranking services and products offered by the gatekeeper itself or by any third party belonging to the same undertaking compared to similar services or products of third party,” DMA art. 6(1)(d).

¹⁴ Multi-homing (DMA art. 5(e)) is the possibility to use the services of more than one platform simultaneously. Interoperability (DMA arts. 5(f)) is both the compatibility of protocols (protocol interoperability) and the possibility to access data in real-time for both the data subject and entities acting on the data subject’s behalf (data interoperability). See JACQUES CRÉMER, YVES-ALEXANDRE DE MONTJOYE & HEIKE SCHWEITZER, COMPETITION POLICY FOR THE DIGITAL ERA 83-84 (2019), <https://data.europa.eu/doi/10.2763/407537> [hereinafter CRÉMER REPORT].

¹⁵ DMA arts. 7-8.

¹⁶ For information on measuring market power, see CRÉMER REPORT, *supra* note 14, at 48-50.

¹⁷ For instance, there has been a vivid debate on the scope of application of the Digital Services Act and the Digital Markets Act, in particular. For a comment on how to best set the criteria for application of the DMA and how to designate “gatekeepers,” see Damien Geradin, *The Digital Markets Act: How Should Ex Ante Rules Look Like?*, PLATFORM L. BLOG (Oct. 23, 2020), <https://theplatformlaw.blog/2020/10/23/the-digital-markets-act-how-should-ex-ante-rules-look-like>. For a discussion of the relationship between dominance and gatekeeping power, see Thosten Käseberg, *Antitrust 2.0—Governance of Oversight Over Digital Gatekeepers*, KLUWER COMPETITION L. BLOG (Dec. 14, 2020), <http://competitionlawblog.kluwercompetitionlaw.com/2020/12/14/antitrust-2-0-governance-of-oversight-over-digital-gatekeepers>.

In this article, we aim to investigate how diffused is the consensus about shifting towards new tools and concepts of competition law among the stakeholders. Do all stakeholder groups share the same understanding and use of the relevant terms and concepts of the DSA and DMA? Or can we identify different attitudes towards these issues?

This is relevant for several reasons. First, a different understanding of legal terms may jeopardize their application by the stakeholders. Think for instance to the many transparency obligations and disclosure duties that require information to be given in a “clear” and “easy-to-understand” manner. The precise enforcement of such duties utterly depends on the behavior of the drafter (mainly medium and big firms) and users, which in turn should have a common understanding and use of said terms. Second, not having a homogeneous understanding and use of terms may cause big reforms like the ones we are discussing to fail to achieve their goals. For example, defining the scope of application of the new rules involves the clear identification of new terms, like gatekeepers. The same can be said with regard to self-preferencing, an anticompetitive conduct which definition is blurry. Third, on a rule-making level, unshared use and understanding of relevant terms may puzzle the consultation process held by the EC over what became the DSA and DMA.¹⁸ According to the Better Regulation agenda 2021,¹⁹ consultations are key to enhance transparency and participation among stakeholders, thus building consensus around new regulatory intervention.²⁰ If some (new) terms are not consistently used, then can consensus really be reached?

For instance, in assessing the feedback it received, the Commission concluded that all stakeholders demanded new rules for gatekeepers which explicitly prohibit “anti-competitive” practices.²¹ This sounds like a consensus at first, but how can we know all respondents are having the same platforms and practices in mind when the key terms they are using are themselves still up to debate?

While some might think of gatekeepers as dominant platforms in the sense of Treaty on the Functioning of the European Union art. 102 others might set a much lower threshold.²² And while some might think that self-preferencing leads to foreclosure, others might view it as necessary to penetrate the market.

¹⁸ The DSA and DMA were developed from three different Inception Impact Assessment documents: EUR. COMM'N, DIGITAL SERVICES ACT PACKAGE: DEEPENING THE INTERNAL MARKET AND CLARIFYING RESPONSIBILITIES FOR DIGITAL SERVICES, INCEPTION IMPACT ASSESSMENT (2020), available at https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12417-Digital-Services-Act-deepening-the-internal-market-and-clarifying-responsibilities-for-digital-services_en; EUR. COMM'N, DIGITAL SERVICES ACT PACKAGE: EX ANTE REGULATORY INSTRUMENT FOR LARGE ONLINE PLATFORMS WITH SIGNIFICANT NETWORK EFFECTS ACTING AS GATE-KEEPERS IN THE EUROPEAN UNION'S INTERNAL MARKET, INCEPTION IMPACT ASSESSMENT (2020), available at <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12418-Digital-Services-Act-package-ex-ante-regulatory-instrument-of-very-large-online-platforms-acting-as-gatekeepers>; and EUR. COMM'N, NEW COMPETITION TOOL ('NCT'), INCEPTION IMPACT ASSESSMENT (2020), available at <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12416-New-competition-tool>.

¹⁹ https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/better-regulation-why-and-how_en.

²⁰ Claudio M. Radaelli, Regulating Rule-making via Impact Assessment, 23(1) *Governance*, 89 (2010).

²¹ See DMA proposal, 7–8.

²² Dominance has been defined as “a position of economic strength enjoyed by an undertaking, which enables it to prevent effective competition being maintained on a relevant market”. While the assessment of market power is based on several factors, an undertaking is generally not considered dominant if it has a market share of below 40% in the relevant market. EUR. COMM'N, COMMUNICATION

In order to investigate these issues, we leverage the power of computational tools. More specifically, we use supervised and unsupervised Machine Learning to analyze the debate preceding the publication of the DSA and DMA proposals and address the following questions: (1) Does the use and understanding of terms related to competition in digital markets differ across different groups of stakeholders? (2) What are stakeholders' attitudes towards certain contentious terms?

On top of these substantive questions, we further discuss (3) the extent to which computational tools can help automate and enrich the analysis of documents that are used as inputs in the rulemaking process. If computational tools really can help in this analysis, then such approaches should be recognized as valuable additions to the "manual" qualitative analysis the Commission currently undertakes.

This Article is structured as follows. The following Part I sets the larger context by outlining some novel competition challenges posed by digital markets that led to the proposed ex-ante regulatory response set out in the DSA and DMA. In Part II, we present our methodology and show, first, that similar opinions are expressed by groups usually belonging to different clusters (i.e., medium and big organizations), and second, groups of stakeholders use central terms of the DSA and DMA in different ways. Lastly, in Part III, we conclude by suggesting what this evidence should tell us about the two proposals and, on a more overarching tone, how computational methods could support EU targeting rules, although very cautiously.

I. COMPETITION IN THE DIGITAL ERA AND THE PROPOSED REGULATORY RESPONSE IN THE DSA AND DMA

There are manifold challenges arising from digital markets that standard competition rules seem not to tackle adequately. This Section will give a snapshot of those inadequacies that were most debated, highlighting what ex-ante rules received the greatest support, as a consequence (A). We will then have a closer look at the DSA and DMA proposals, to see in what way such debate and the perceived needs translated into actual norms (B).

A – Why We Need the DSA and DMA: New Buzzwords

Consumers benefit in many ways from the impressive development of digital markets.²³ However, lawmakers and scholars alike have been emphasizing that digital markets show certain characteristics which are likely to favor highly concentrated markets.

FROM THE COMMISSION — GUIDANCE ON THE COMMISSION'S ENFORCEMENT PRIORITIES IN APPLYING ARTICLE 82 OF THE EC TREATY TO ABUSIVE EXCLUSIONARY CONDUCT BY DOMINANT UNDERTAKINGS, 2009/C 45/02, at 8-9.

²³ See DSA recital 1. On the DSA see Alexandre de Stree & Sally Broughton Micova, CENTRE ON REGULATION IN EUROPE, DIGITAL SERVICES ACT – DEEPENING THE INTERNAL MARKET AND CLARIFYING RESPONSIBILITIES FOR DIGITAL SERVICES at 39 (2020), <https://cerre.eu/publications/digital-services-act-responsibility-platforms/> (also referred as CERRE DSA report, in the following).

1. Seizing Market Power of Digital Platforms

First, many business models in the digital realm are characterized by strong returns to scale.²⁴ Second, incumbents in online markets are particularly hard to dislodge due to substantial network effects.²⁵ Third, due to the data dependency of many online services, established players might hold a competitive edge over small contestants by leveraging the power of the large amounts of data they accumulate.²⁶

While the extent of these advantages could be limited by multi-homing and interoperability,²⁷ there still is a very real chance that certain platforms accumulate some kind of “gatekeeping” power and impose the prices, conditions, and level of transparency they deem appropriate, for their own convenience. In its investigation report on competition in digital markets, the US Congress Subcommittee on Antitrust, Commercial Law and Administrative Law addressed the last-mentioned aspect: “Without transparency or effective choice, dominant firms may impose terms of service with weak privacy protections that are designed to restrict consumer choice, creating a race to the bottom.”²⁸ This would further increase the danger of “tipped markets” in which one company takes the large majority of the market share.²⁹

While there is broad consensus that these dynamics can be highly problematic, the exact threshold for a tipped market, or the precise definition of a “gatekeeper” position arising from such a situation is still up to debate.³⁰ Especially the relation between the legal concept of dominance and the definition of gatekeepers has been the subject of a vivid debate among scholars.³¹

²⁴ CRÉMER REPORT, *supra* note 14, at 3; Michal S. Gal & Nicolas Petit, *Radical Restorative Remedies for Digital Markets*, 37 BERKELEY TECH. L. J. (forthcoming 2021) (manuscript 5-6), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3687604#; STIGLER CENTER FOR THE STUDY OF THE ECONOMY AND THE STATE, REPORT OF THE COMMITTEE FOR THE STUDY OF DIGITAL PLATFORMS, MARKET STRUCTURE AND ANTITRUST SUBCOMMITTEE 14 (2019) [hereinafter STIGLER REPORT]; Org. for Econ. Coop. & Dev. [OECD], *Executive Summary of the Roundtable on Algorithms and Collusion*, at 5, OECD Doc. DAF/COMP/M(2017)1/ANN3/FINAL (Sept. 26, 2018), [https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DAF/COMP/M\(2017\)1/ANN3/FINAL&docLanguage=En](https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DAF/COMP/M(2017)1/ANN3/FINAL&docLanguage=En).

²⁵ CRÉMER REPORT, *supra* note 14, at 3; Peter Georg Picht & Gaspare Tazio Loderer, *Framing Algorithms: Competition Law and (Other) Regulatory Tools*, 42 WORLD COMPETITION 391, 406 (2019).

²⁶ CRÉMER REPORT, *supra* note 14, at 2.

²⁷ When users can multi-home, they might switch to better service providers or use services in parallel, which will increase competition. Interoperability often is a pre-condition for multi-homing and might further allow users to unbundle or establish complementary services. Cr mer report, *supra* note 14), at 23, 83-84. For definitions of multi-homing and interoperability, see note 11.

²⁸ SUBCOMM. ON ANTITRUST, COM. & ADMIN. L. OF THE COMM. ON THE JUDICIARY, H. COMM. ON THE JUDICIARY, 116TH CONG., INVESTIGATION OF COMPETITION IN DIGITAL MARKETS. (2020). The Subcommittee report also mentions manipulative design interfaces, so-called dark patterns, nudging consumers into certain choices. *Id.* at 53.

²⁹ DIGIT. COMPETITION EXPERT PANEL, HM TREASURY, UNLOCKING DIGITAL COMPETITION: REPORT OF THE DIGITAL COMPETITION EXPERT PANEL 3-4, 88 (2019), <https://doi.org/10.17639/wjcs-jc14>.

³⁰ Discussions about how to best designate “gatekeepers” started before the proposal was published and continued in the aftermath of its publication. For a critical view on the gatekeeper definition of Art. 3 DMA, see Lu s Cabral, Justus Haucap, Geoffrey Parker, Georgios Petropoulos, Tommaso Valletti, & Marshall Van Alstyne, Eur. Comm’n, *The EU Digital Markets Act: A Report from a Panel of Economic Experts*, at 9 (2021), <https://publications.jrc.ec.europa.eu/repository/handle/JRC122910>; Cristina Caffarra & Fiona Scott Morton, *The European Commission Digital Markets Act: A Translation*, VOX EU CEPR (05 Jan. 2021), <https://voxeu.org/article/european-commission-digital-markets-act-translation> (accessed 3 Apr. 2021).

³¹ For a comprehensive discussion of possible criteria for gatekeepers, see ALEXANDRE DE STREEL, CTR. ON REGUL. IN EUR., DIGITAL MARKETS ACT: MARKING ECONOMIC REGULATION OF PLATFORMS FIT FOR THE DIGITAL AGE 35-44 (24 Nov. 2020) [hereinafter CERRE DMA report]. While UK authorities intentionally eschewed the traditional notion of dominance (see, e.g., HM Treasury, ‘Unlocking digital competition, Report of the Digital Competition Expert Panel’, 13 March 2019,

2. Tackling Anti-Competitive Conduct

Besides these structural challenges, a lot of the competition law debate on digital markets has centered around some specific anti-competitive conducts. First, the issue of *algorithmic* collusion and *tacit* collusion was heavily discussed. Art. 101 TFEU notoriously only condemns collusive arrangements, leaving uncoordinated parallel behavior untouched. For a long time, this was not perceived as a problem since tacit collusion is almost never stable on non-digital markets.³² However, the increased concentration, transparency, entry barriers and interaction frequency of online markets make tacit collusion, especially through algorithmic means, a theoretically very plausible scenario.³³ Although the practical relevance of this phenomenon is still disputed,³⁴ there is no denying that collusion has received considerable attention among scholars and competition authorities.

Second, the list of digital competition buzzwords further includes “self-preferencing.”³⁵ Whenever platforms somehow influence consumer choice, e.g., by presenting offers from business customers in a certain order, they have the possibility to favor their own products or services. The *Google Shopping* case established that self-preferential placements are, indeed, not compatible with competition law.³⁶ Only if the parameters used to rank products are transparent,³⁷ will it be possible to know whether an online platform is distorting competition by preferencing certain offers, leaving consumers in the dark about the “trade-offs they are facing,” and hence inhibiting competition in a significant manner.

3. Regulatory Reform: The Need For Ex-Ante Rules

Existing competition and consumer protection laws often fall short in addressing these issues, both in the EU and the United States. For instance, the US Subcommittee on Antitrust has noted that “some of these business practices are a detriment to fair competition, but they do not easily fit the existing categories identified by the Sherman Act, namely ‘monopolization’ or ‘restraint of trade’ or

<https://www.gov.uk/government/publications/unlocking-digital-competition-report-of-the-digital-competition-expert-panel> (accessed 25 February 2021), at 55, Germany’s tenth amendment of its Act Against Restraints of Competition (Gesetz gegen Wettbewerbsbeschränkungen, GWB) explicitly includes dominance in one or several markets as part of its definition of “paramount significance” (§ 19(1) no. 1 GWB), the equivalent to the DMA’s “gatekeepers.”

³² Francisco Beneke & Mark-Oliver Mackenrodt, *Remedies for Algorithmic Tacit Collusion*, 9 J. ANTITRUST ENFT, 152, 158-159 (2021).

³³ BUNDESKARTELLAMT & AUTORITE DE LA CONCURRENCE, ALGORITHMS AND COMPETITION, at II (2019), https://www.bundeskartellamt.de/SharedDocs/Publikation/EN/Berichte/Algorithms_and_Competition_Working-Paper.html;jsessionid=BCD024895C5890BEDCAAAB208FBCC6B8F.2_cid387?nn=3591568.

³⁴ For a critical view, see Salil K. Mehra, *Robo-Seller Prosecutions and Antitrust’s Error-Cost Framework*, COMPETITION POL’Y INT’L: ANTITRUST CHRON. (15 May 2017), <https://www.competitionpolicyinternational.com/robo-seller-prosecutions-and-antitrusts-error-cost-framework/> (accessed 3 Apr. 2021).

³⁵ For an in-depth discussion, see Damien Geradin & Dimitrios Katsifis, “Trust Me, I’m Fair”: *Analysing Google’s Latest Practices in Ad Tech from the Perspective of EU Competition Law*. 16 EUR. COMPETITION J., II (2020). Self-preferencing was at the heart of the Microsoft saga (on which, see John P. Jennings, Comparing the US and EU Microsoft Antitrust Prosecutions: How Level Is the Playing Field?, 2 *Erasmus L. & ECON. REV.* 71 (2006)).

³⁶ In the *Google Shopping* case, self-preferential placements were deemed not compatible with competition law: Commission Decision C(2017) 4444 of 27 June 2017, *Google Search (Shopping)*, 2018 O.J. (C 9) ¶¶ 9-10 of summary decision. See Pinar Ackman, *The Theory of Abuse in Google Search: A Positive and Normative Assessment Under EU Competition Law*, 2017 J. L., TECH. & POL’Y 301 (2017).

³⁷ Picht and Loderer, *supra* note 18, at 416.

the Clayton Act.³⁸ The American Choice and Innovation Online Act of June 11, 2021³⁹, can thus be read as a means to prohibit the usage of exploitative practices by large online platforms and to promote users.

In the debate surrounding the DSA and DMA proposals, general shortcomings of EU competition rules when dealing with opaque online platforms have been highlighted.⁴⁰ Relying exclusively on Arts. 101 and 102 TFEU might mean that the Commission could only act on a case-by-case basis, for a very limited set of platforms, ex post, and only after lengthy investigations.⁴¹

Given that Section 2 of the Sherman Act (15 U.S.C. § 2) and Section 7 of the Clayton Act (15 U.S.C. § 18) are even more narrow than comparable EU competition norms, their suitability to achieve a satisfactory level of fairness and competition might be more limited, especially regarding effective remedies.⁴²

³⁸ SUBCOMM. ON ANTITRUST, *supra* note 9, at 396.

³⁹ US House Judiciary Committee, *supra* note 10. See e.g. Sect. 2.3 of the Act prohibiting discriminatory conduct against business users by those operating 'covered platforms' (a concept largely assimilable to the European 'gatekeeper').

⁴⁰ The Crémer report points out several criticalities: (1) Not all gatekeepers enjoy a dominant position in the sense of Art. 102 TFEU; (2) the relevant market might be substantially harder to define than in non-digital cases; (3) not every problematic practice has a demonstrable effect on the relevant market. The authors conclude that greater emphasis should be put on the theory of harm, instead. Crémer report, *supra* note 14, at 3-4. Moreover, digital markets are often moving at a rapid pace, which is not necessarily a characteristic they share with competition law. Hence, there are concerns whether competition law could be applied with the necessary speed to address urgent competition needs. CERRE DMA report, *supra* note 24, at 59; Recital 5 DMA

⁴¹ This is not to say that existing competition norms are generally useless to address novel competition issues; in fact, EU case law has shown the opposite. *E.g.*, Google Search (Shopping) Case C(2017) 4444, 27 June 2017; Commission Decision C(2018) 4761 of 18 July 2018., Google Android, 2019 O.J. (C 402). On the national level the German competition authority has taken action against certain data collection practices of Facebook (see Bundeskartellamt [Federal Cartel Office] 6 Feb. 2019, B6-22/16, Bundeskartellamt, at 6, https://www.bundeskartellamt.de/SharedDocs/Entscheidung/EN/Entscheidungen/Missbrauchsaufsicht/2019/B6-22-16.pdf?__blob=publicationFile&v=5). Also with a focus on data collection, the Commission has opened an investigation against Amazon in 2019 (European Commission Press Release IP/19/4291, Antitrust: Commission Opens Investigation into Possible Anti-Competitive Conduct of Amazon (17 July 2019), https://ec.europa.eu/commission/presscorner/detail/en/ip_19_4291). For older cases, see Spencer Weber Waller, *Access and Information Remedies in High-Tech Antitrust*, 8 J. COMPETITION L. & ECON. 575, 576 (2012). Regarding the United States, some argue in a similar vein: "This is not to say that the use of the antitrust laws should be abandoned. If history is a guide, there is a meaningful possibility that antitrust enforcement activities will produce value commensurate with their costs." TOM WHEELER, PHIL VERVEER, & GENE KIMMELMAN, SHORENSTEIN CTR. ON MEDIA, POL. & PUB. POL'Y, HARVARD KENNEDY SCH., NEW DIGITAL REALITIES; NEW OVERSIGHT SOLUTIONS IN THE U.S. 26 (2020), https://shorensteincenter.org/wp-content/uploads/2020/08/New-Digital-Realities_August-2020.pdf.

⁴² Wheeler et al., *supra* note 33, at 24-26. On fairness in competition law see Ariel Ezrachi (2018) EU Competition Law Goals and the Digital Economy, *Oxford Legal Studies Research Paper* No. 17/2018 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3191766) (accessed 27 January 2021), at 16; Maurice Stucke & Ariel Ezrachi (2020) COMPETITION OVERDOSE: HOW FREE MARKET MYTHOLOGY TRANSFORMED US FROM CITIZEN KINGS TO MARKET SERVANTS (Harper Business); Shoshana Zuboff (2019) THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER (PublicAffairs).

In light of these interconnected challenges for consumer protection and competition, a consensus has been reached on the need for new ex-ante rules⁴³ to complement the existing legal framework.⁴⁴

A – How the DMA and DSA Proposals Respond to the Quest for New Pro-Competitive Rules

The European Commission’s vision of these rules was first outlined in three separate inception impact assessments,⁴⁵ that were open to consultation by stakeholders. As a result, the DMA and DSA normative proposals were adopted that encapsulate such discussion. In the following, we will briefly present their content.

1. Defining The Scope: The Gatekeepers

The DSA applies to all “intermediary services,” while the scope of the DMA is limited to “core platform services” offered by “gatekeepers” as defined in Article 3 of the DMA. The DSA replicates the GDPR’s approach to applicability, hence applying to all services provided to EU citizens “irrespective of the place of establishment of the providers of those services.”⁴⁶ This is a characteristic it shares with the DMA.⁴⁷

However, the DMA is much more limited in scope since its goal is not to lay down a fundamental framework, but rather to complement existing competition law norms with respect to a very specific set of market players (the gatekeepers).⁴⁸ Regarding its *ratione materiae*, the DMA has a more limited scope (the core platform services listed in Article 2 Section 2),⁴⁹ to be in line with its much more specific objective of “ensuring contestable and fair markets ... where gatekeepers are present.”⁵⁰ Given the centrality of the “gatekeeper” concept, different understandings of this term could cast doubts at whether all stakeholders mean the same when they express their support for additional rules for “gatekeepers.”

⁴³ Discussions on ex-ante rules to complement ex post antitrust enforcement in digital markets can be found in CERRE DMA report (n. 31), at 26; Luis Cabral; Justus Haucap; Geoffrey Parker; Georgios Petropoulos; Tommaso Valletti; & Marshall Van Alstyne, The EU Digital Markets Act: A Report from a Panel of Economic Experts, JRC122910, 9 February 2021 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3783436# (accessed 1 July 2021); Jonathan B. Baker, Protecting and Fostering Online Platform Competition: The Role of Antitrust Law, 17(2), *J. of Competition Law & Economics*, 493 (2021) (discussing the role of regulation to supplement antitrust law in fostering competition in online platforms).

⁴⁴ See, e.g., ALGORITHM WATCH, GOVERNING PLATFORMS – FINAL RECOMMENDATIONS (2020), https://algorithmwatch.org/wp-content/uploads/2020/10/Governing-Platforms_DSA-Recommendations.pdf, at 1; CERRE DSA report (*supra* note 23) at 39; Comm. on the Internal Mkt. & Consumer Prot., Eur. Parliament, *Draft Report with Recommendations to the Commission on Digital Services Act: Improving the Functioning of the Single Market*, para. 15, at 5, 2020/2018(INL) (24 Apr. 2020); Comm. on Legal Affairs, *Draft Report with recommendations to the Commission on a Digital Services Act: adapting commercial and civil law rules for commercial entities operating online*, para. 2, at 5, 2020/2019(INL) (22 Apr. 2020).

⁴⁵ See *supra* note 16.

⁴⁶ DSA art. 1, § 3.

⁴⁷ DMA art. 1, § 2.

⁴⁸ See *supra* notes 7-11 and accompanying text.

⁴⁹ Note that electronic communication network and services markets are exempted from the scope of the proposal (DMA art. 1 §3).

⁵⁰ DMA art. 1 §1 (emphasis added).

2. Regulating Conducts: New Ex-Ante Rules

Turning to some of the substantive rules for said gatekeepers, under the new law gatekeepers will have to grant data access to business users for the data they generated on the platform and allow for multi-homing and interoperability.⁵¹ Self-preferencing will be explicitly prohibited,⁵² as will preventing users from uninstalling certain pre-installed tools.⁵³ With a view to advertisement markets, Article 5 Section g of the DMA would oblige gatekeepers to provide information about pricing and performance measuring tools to consumers of core platform services within the meaning of Article 3 Section 7 of the DMA.⁵⁴ This would allow consumers to assess how satisfied they are with the advertisement product they are paying for.⁵⁵

With a view to the connection between transparency and competition mentioned above, a look at the DSA is helpful to complete the picture. Its Articles 12(1), 13, 23, 24, 29, and 33 establish comprehensive but differentiated disclosure and reporting duties regarding ranking, advertisement, and content moderation practices.⁵⁶ The more pronounced transparency obligations for very large online platforms within the meaning of Article 24 of the DSA reflect the differentiated approach the Commission took for the design of the DSA, explicitly mentioned in Recital 39 of the proposal.

To sum up, this section has shown that certain characteristics of digital services might impair competition. It has also sketched out the debate on why existing rules might be insufficient to prevent this and how the DMA and DSA proposals of the European Commission seek to change this.

II. A COMPUTATIONAL ANALYSIS OF THE DSA AND DMA CONSULTATION PROCESS

In motivating its twin proposals, the Commission reports that the “vast majority of respondents” in the consultation process “considered that dedicated rules on platforms should include prohibitions and obligations for gatekeeper platforms.”⁵⁷ More specifically, in the DMA this majority believed that the “the proposed list of problematic practices, or ‘blacklist,’ should be targeted to clearly unfair and harmful practices of gatekeeper platforms.”⁵⁸ In the DSA, the quest for “algorithmic accountability and transparency audits, especially with regard to how information

⁵¹ DMA art. 5, §§ f, h, g.

⁵² DMA art. 6, §1(d).

⁵³ DMA art. 5, §b.

⁵⁴ DMA art. 5, §g. *See also* DMA art. 3, §7.

⁵⁵ DMA art. 6, §1(g). While these obligations are rather specific, Article 10 of the DMA would open the door to add further duties in the future if a market investigation pursuant to Article 17 of the DMA identified a need to do so for the sake of safeguarding fair competition.

⁵⁶ Just to name those duties for very large online platforms within the meaning of Article 25 of the DSA. Article 29 of the DSA would entail the obligation to provide information on the use of recommender systems and their parameters. Article 33 sets out comprehensive transparency obligations for very large online platforms.

⁵⁷ DMA, Explanatory Memorandum at 8 (summarizing the results of stakeholder consultations and impact assessments).

⁵⁸ *Ibid.*

is prioritized and targeted” online comes from “a wide category of stakeholders,” and is particularly voiced by “civil society and academics.”⁵⁹

In this chapter, we ask whether these ex-ante rules for very large platforms and gatekeepers are what stakeholders demanded in the consultation process, and whether their actual wording in the DSA and DMA proposals reflects the way each stakeholder group uses the relevant terms. This is a relevant step, as it is important that the addressees of such duties (typically digital firms) and the beneficiaries (individuals, micro and small organizations using platforms) agree on their meaning.

To do so, we use computational text analysis techniques to analyze the replies to the questionnaires and position papers submitted by stakeholders to the EU consultation process for both proposals.

A. *Data and Methodology*

To analyze these stakeholder documents, we created a special scraper algorithm, which allowed us to download all the files automatically, convert them into text, and split them into three clusters. In doing this, we relied on the Commission’s categorization, based on the organization size of the feedback contributors.⁶⁰ We then aggregated the different sub-categories into three corpuses, based on type and organizational size of the feedback contributor:

Corpus A (individuals and micro-organizations),
Corpus B (small companies/organizations), and
Corpus C (medium and large companies/organizations).

This clustering scheme was informed by a preliminary analysis of our data, which is explained in more detail in the Appendix (Annex 1). Interestingly and importantly, it showed that the classical “small and medium enterprises” group was not equally applicable to our data, since the feedback contributors do not only comprise enterprises, but a more diverse set of actors.

To discern differences in the use of certain key terms across stakeholder groups (i.e., a different semantic understanding of identical terms), we leveraged Word Embedding Models to quantify evidence of such differing understandings. This technique has already been used in various Natural Language Processing tasks,⁶¹

⁵⁹ DSA, explanatory memorandum at 9. See also Algorithm Watch (SUPRA NOTE 46), at 1; CERRE DSA REPORT (supra note 23) at 39; *Draft Report of the Committee on the Internal Market and Consumer Protection with Recommendations to the Commission on Digital Services Act: Improving The Functioning of the Single Market* at 5, 2020/2018(INL) (n 39); Eur. Comm’n, *White Paper on Artificial Intelligence - A European approach to excellence and trust*, at 15, COM(2020) 65 final

⁶⁰ The Commission distinguishes between (1) individuals, micro (< 10 employees), (2) small (< 50 employees), (3) medium (< 250 employees), and (4) large (250 or more) organizations as well as between different types of feedback contributors (i.e., in the DSA, p. 8 you find that respondents are: the general public (66%), companies/businesses organizations (7.4%), business associations (6%), and NGOs (5.6%) authorities (2.2%), academic/research institutions (1.2%), trade unions (0.9%), and consumer/environmental organizations (0.4%)). DSA, Explanatory Memorandum at 8.

⁶¹ For instance, word embeddings have been used by sociologists to investigate the meaning of the term ‘class’, to predict conflict, or to classify documents. Austin C. Kozlowski, Matt Taddy & James A. Evans, *The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings*, 84 AMERICAN

and recently also in the Computational Law literature. More specifically, it has been demonstrated to be very powerful and useful in providing insights into latent differences in how language is used in legal scholarship⁶² and is starting to be discussed in the computational antitrust literature.⁶³

The core of this technique consists in training a special neural network to convert each word contained in a corpus of texts into a vector, i.e., a set of numbers.⁶⁴ While a simple algorithm would require formulating explicit rules to somehow approximate the semantic meanings of words, ML (or the neural network, to be precise) learns the implicit rules directly from the data we feed it. The resulting vectors are based on the frequency with which words occur next to each other. The neural network tracks their relative positions in each phrase of the corpus and the correlation between words. The stronger two words are correlated (in their occurrence – and so in their semantic meaning)⁶⁵ in the corpus the model was trained in, the closer the corresponding vectors will be located to each other.

However, models trained on different corpuses are not directly comparable as they depend on the corpus the model was trained on. The vector of a single word alone does not provide meaningful insights. To test if there is evidence of a different semantic use of the same words between two texts, we assess the distance between vectors from two different corpuses corresponding to the same words. To align them, we transform the two models geometrically.⁶⁶ This allows to understand how a vector in one corpus relates to the vector of another corpus. After the transformation, the vectors of the two aligned corpuses are comparable. Therefore, for each corpus we trained a different word embedded space, and we aligned each pair of words occurring in both corpuses through the means of Unsupervised Vector Space Alignment.⁶⁷

SOCIOLOGICAL REVIEW 905 (2019); Andrey Kutuzov, Erik Velldal & Lilja Øvrelid, *Tracing Armed Conflicts With Diachronic Word Embedding Models* in PROCEEDINGS OF THE EVENTS AND STORIES IN THE NEWS WORKSHOP 31–36 (Tommaso Caselli et al., eds.); Xiao Yang, Craig Macdonald & Iadh Ounis, *Using Word Embeddings in Twitter Election Classification*, 21 INFORMATION RETRIEVAL JOURNAL 183 (2018).

⁶² Julian Nyarko & Sarath Sanga, *A Statistical Test for Legal Interpretation: Theory and Applications* (Nov. 25 2020) (unpublished article) (on author's website) https://juliannyarko.com/wp-content/uploads/other/nyarko_sanga_legal_interpretation.pdf.

⁶³ For a discussion of computational techniques in antitrust law and their implications, see the following: Thibault Schrepel, *Computational Antitrust: An Introduction and Research Agenda*, 1 STAN. COMPUTATIONAL ANTITRUST 1 (2021); Giovanna Massarotto & Ashwin Ittoo, *Gleaning Insight from Antitrust Cases Using Machine Learning*, 1 STAN. COMPUTATIONAL ANTITRUST 16.

⁶⁴ The Neural Network used in this research in particular is a LSTM (Long-Short Term Memory Network). See Sepp Hochreiter & Jürgen Schmidhuber, *Long Short-term Memory*, 9 *Neural Computation* 1735 (1997).

⁶⁵ This is based on the “distributional hypothesis,” which assumes that words which frequently occur together are usually also semantically related. While this approach might seem too simple to capture complex semantic meanings, the success of algorithms relying on it suggests that the claim has some merit. Edgar Altszyler, Mariano Sigman, Sidarta Ribeiro, & Diego Fernandez Slezak, *Comparative Study of LSA vs Word2vec Embeddings in Small Corpora: A Case Study in Dreams Database*, 56 *CONSCIOUSNESS AND COGNITION* 178.

⁶⁶ To perform this transformation, we used a ‘control vocabulary’, containing a list of words that we can safely assume that share the same semantical meaning. The list of 1,189 words we used is, in fact, composed mainly of numbers and stop-words (like e.g. ‘the’). We are thankful to Professor Julian Nyarko from Stanford University for providing us with a first list of control keywords, to which we further added almost 2,000 stop-words and numerals we took from the different corpuses. We manually selected our control vocabulary. We used the glossaries contained in all EU directives and regulations recalled in the DSA and DMA proposals and published in the OJUE. Further, we manually coded the questionnaires (used in the consultation) and selected terms of interest.

⁶⁷ We used a special algorithm provided by Facebook in the library FastText. (<https://github.com/facebookresearch/fastText>), used in Python. Piotr Bojanowski, Edouard Grave,

Having aligned our three corpuses, we are able to compute the distance between the same terms from different corpuses. However, how do we know that the distances we find are not just random, but actually based on substantial uses and understandings between stakeholders? To see if there is evidence for a statistically significant semantic difference between the use of a term between the different stakeholder groups, we must perform a statistical test (see Appendix, Annex 2 for detailed description). Otherwise, we would not be able to tell whether the differences we find between corpuses are actually relevant or merely signs of random, non-semantic differences between our corpuses. By modeling the theoretical distribution of non-semantic differences for the terms in our corpus, we can compare the distance we would expect to see if there was no semantic difference with the distance we observe. In this way, we can conclude with a certain level of confidence that the distance we observe between our corpuses is more than just random or syntactical.

In order to gain deeper insights into possible reasons for a semantic difference in the use of key words between different corpuses, we also leveraged the tool of Sentiment Analysis,⁶⁸ applied to sentences of the two compared corpuses where the specific key word appears. Sentiment analysis is a Natural Language Processing technique, which classifies a sentence, or a paragraph based on the use of specific words and their location inside the text, giving as a score a value of the positive or negative sentiment inside the particular text. These two values are aggregated to a compound value, which gives a score to the overall sentiment ranging from -1 (totally negative) to +1 (totally positive).

B. Results and Discussion: Different Groups, Different Uses?

With these tools at hand, we were able to find a significant difference for 1,865 word pairs between Corpuses A and C. Between Corpuses A and B we found 2,184 statistically significant differences and 1,113 between stakeholder groups B and C.⁶⁹

In the following, we will only discuss the most interesting differences we found, that are relevant to (1) key actors and structure of digital markets, (2) their anticompetitive conduct, and (3) the identified remedies and ex-ante rules. The results are summarized in Tables 1 to 3 respectively.

Summary of results

Table 1: Key actors and market structures

Term	Distance AB	Distance BC	Distance AC	Close words A	Close words B	Close words C
Dominant	0.515 (0.65)	0.797 (0.50)	1.530 (0.02)**	Self-preferencing		Policymakers, quasi-judicial

Armand Joulin & Tomas Mikolov, Enriching Word Vectors with Subword Information (June 19 2017) (unpublished manuscript) (on file at <http://arxiv.org/abs/1607.04606>) (accessed 22 January 2021).

⁶⁸ Firoz Khan et al., *Sentiment Analysis of Twitter Data*, 6 INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH 15 (2018).

⁶⁹ Note that many of these words are not of particular interest for us as they might be specific to a position paper of a certain company (e.g. 'Gmail' in Google's submissions). However, some of the key buzzwords surrounding competition law and new ex-ante remedies show statistically significant differences.

Gatekeepers	1.064 (0.22)	1.367 (0.11)	1.486 (0.03)**	Content		Unsatisfactory
Monopolistic	1.023 (0.25)	1.473 (0.06)*	1.631 (0.00)**	Discouraging, profitability	Vulnerability, linking	Higher-cost, welfare
Monopolization*	1.323 (0.109)	1.432 (0.078)	1.609 (0.00)**	Endangers, non- dominant, data		Operations
Newcomers	1.469 (0.03)**	1.571 (0.04)**	1.192 (0.18)	Non-existent, none	Tech	Start-ups, destroyed

**Monopolization" was used a relatively small number of times (only 285), compared to other reported words. We decided to include it to show the semantic distance with neighboring concepts.

Table 2: Anticompetitive conduct

Term	Distance AB	Distance BC	Distance AC	Close words A	Close words B	Close words C
Abusive	1.216 (0.14)	1.162 (0.20)	1.543 (0.04)**	Competing		Tacit, monopoly
Collusion	1.590 (0.01)***	1.420 (0.08)*	1.396 (0.08)*	policing	tacit, data	
Coordinated	1.395 (0.06)*	1.110 (0.23)	1.541 (0.02)**	Cooperation, dialogue	Tacit, legitimate	Future-proof, EU-level
Pro-competitive	1.445 (0.04)**	1.295 (0.13)	1.022 (0.29)	comparison	user-friendly	
Ranking	1.182 (0.15)	1.644 (0.02)**	1.452 (0.05)**		guidelines, improve, oversight	appearance, disclosing
Self-favoring	.	.	1.589 (0.00)**	combating		debates
Self-preferencing	1.393 (0.04)**	1.457 (0.07)*	1.190 (0.18)	monopolizing	over- regulation, reports	
Tacit	1.439 (0.04)**	0.923 (0.40)	1.274 (0.13)		Threatens	
Tipping	1.509 (0.03)**	1.584 (0.03)**	1.213 (0.17)		bottleneck, nudge	unwanted
Uncompetitive	1.478 (0.03)**	1.199 (0.18)	1.499 (0.03)**	single-sided, state-run	EU- commission, institutions	

Table 3: Remedies and ex-ante antitrust reform

Term	Distance AB	Distance BC	Distance AC	Close words A	Close words B	Close words C
Blacklist	1.402 (0.05)	1.063 (0.27)	1.580 (0.01)**	functionalizati on		Dominance- based, problem
Data-sharing	1.566 (0.02)**	.	.	recycled	differentiation	
Disproportionally	1.590 (0.01)**	1.044 (0.28)	1.154 (0.20)	Public-interest, bans	Ensure, competition, incentives	
Interoperability	1.665 (0.01)**	1.150 (0.21)	0.756 (0.49)	reliability, trustworthy	Licensing	
Overregulated	1.500 (0.03)**	.	.	Sellers	Tax-like	

Pro-competitive	1.445 (0.04)**	1.295 (0.13)	1.022 (0.29)	comparison	user-friendly	
Underenforcement	1.463 (0.03)**	1.283 (0.13)	1.513 (0.02)**	Complaints, unbureaucratic	Consensus	Misconceptions , improvements
Unregulated	1.361 (0.07)*	1.566 (0.04)**	1.822 (0.00)***	not-sufficient		mitigation
Welfare	0.989 (0.26)	1.543 (0.04)**	0.626 (0.65)		Economic, rights	mobility

*Note: The "Distance" columns report the distance between the vectors of the same words for each corpus pair with the respective p-value in parentheses. A grey field indicates that a word was not used in both of the respective corpuses. The "Close Words" columns shine a light at some of the concepts that were closely related with the term in question in the corpuses for which there was a statistically significant distance between the terms. The asterisks indicate significance at a 0.001 (***), 0.05 (**), and 0.1 (*) level, respectively.*

Key Actors

Figure 1: Projection of Aligned Word Embedding. Aligned Vector Space Model Corpora

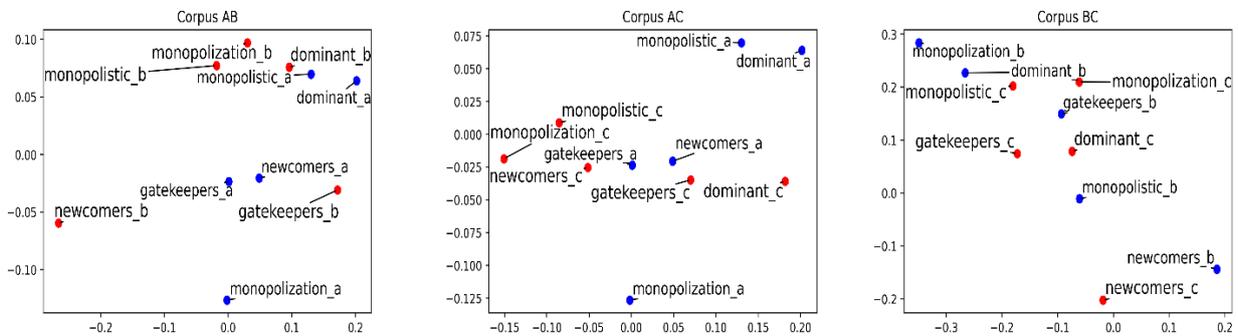
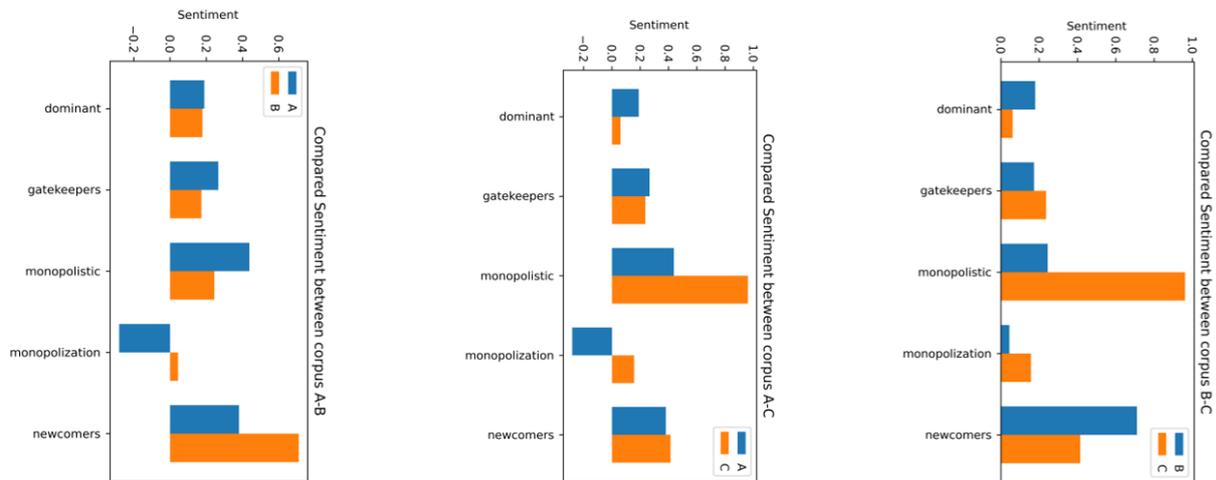


Figure 2: Sentiment Analysis. Compared Corpora Related to Key Actors



Starting with terms related to key actors, and especially the applicability of competition law and the DMA, we see a statistically significant difference between micro companies/organizations and the medium/big companies and organizations, for an absolutely central term: “gatekeepers.” This can be linked back to the Commission’s summary of the consultation, where it pointed out that some stakeholders considered the concept of a gatekeeper too broad.

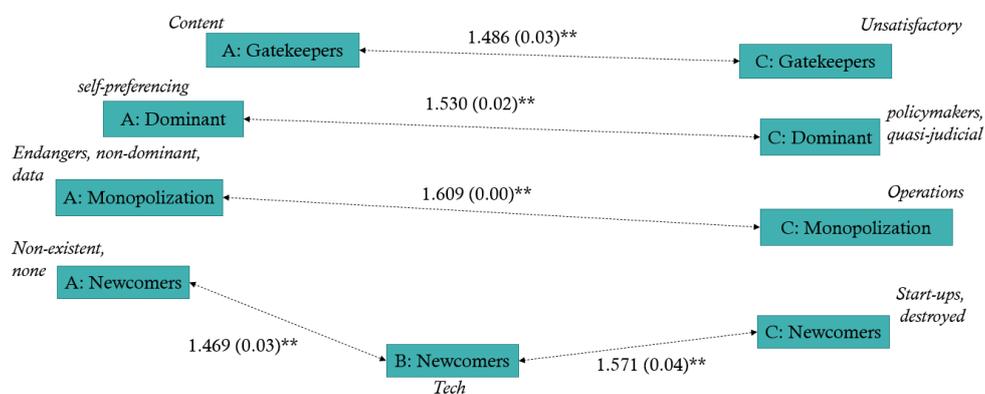
Here, our computational analysis confirms the qualitative assessment of the Commission. The question of how “gatekeeper” should be defined was hotly debated⁷⁰ and it seems like the differing standpoints also translated into a different use of the term “gatekeeper” in the position papers for all three corpora.

Another term that has risen in importance in the recent debates on digital markets is “newcomers,” the counterpart of “gatekeepers.” We find that the term is used differently between both Corpus A and B as well as between Corpus B and C. Interestingly (Fig. 3), it is closely related with “non-existent” in Corpus A and

⁷⁰ See DSA, Explanatory Memorandum, 8.

“destroyed” in Corpus C, which seems to hint at the difficult standing of small tech start-ups on certain digital markets.

Figure 3. Distances and closest words for terms related to “key actors”



Remarkably, we also find a significant difference for the term “dominant,” with Corpus A showing a close connection to the term “self-preferencing.” This is noteworthy with a view to Art. 102 TFEU, which applies only to firms holding a “dominant” position. This finding suggests that not only new competition concepts, but also established ones are perceived differently, in this case by micro-organizations/individuals and medium/big organizations.

Unlike “dominant,” EU competition law does not make use of the term “monopoly” or related concepts. However, they are frequently used in debates on competition law, including those on the DSA and DMA. Our analysis shows that there is no perfect agreement between different stakeholder groups concerning the precise content of terms like “monopolistic” or “monopolization.” While the novelty of a term like “gatekeeper” might to some extent explain differences in its use, it is surprising that more established terms are used in an equally non-uniform way. This could be due to the fact that many respondents are non-Europeans and therefore not used to the EU competition jargon. However, as the close words analysis for “monopolization” reveals, this finding might also be related to the debates on how even non-dominant firms might gain a position that enables them to, e.g., effectively bar start-ups (“newcomers”) from entering the market. The fact that also “data” is closely related to “monopolization” could be interpreted as a hint of how new technologies can be a drive for economic giantism, while introducing uncertainty around relatively well-established terms.

Nevertheless, it needs to be noted that we did not find a significant difference for terms like “platform,” “market power,” or “dominance” which should be encouraging.

Anti-Competitive Conduct

Figure 4: Projection of Aligned Word Embedding. Aligned Vector Space Model Corpora

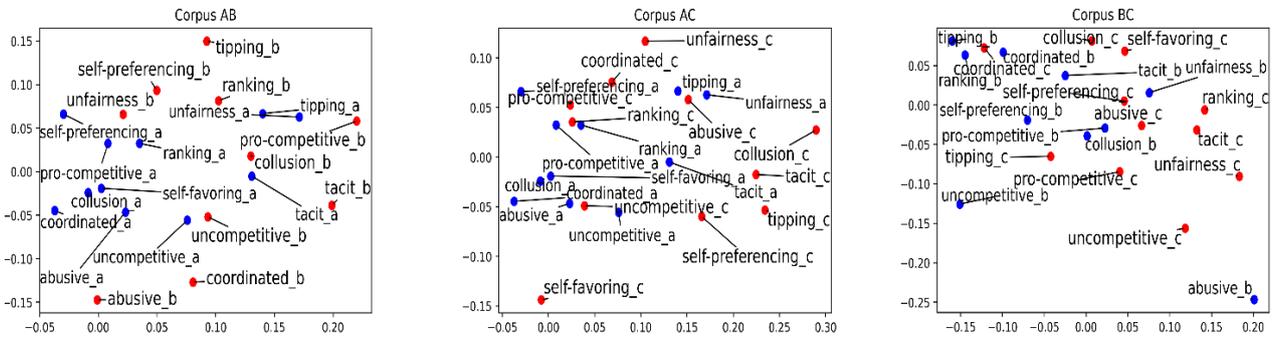
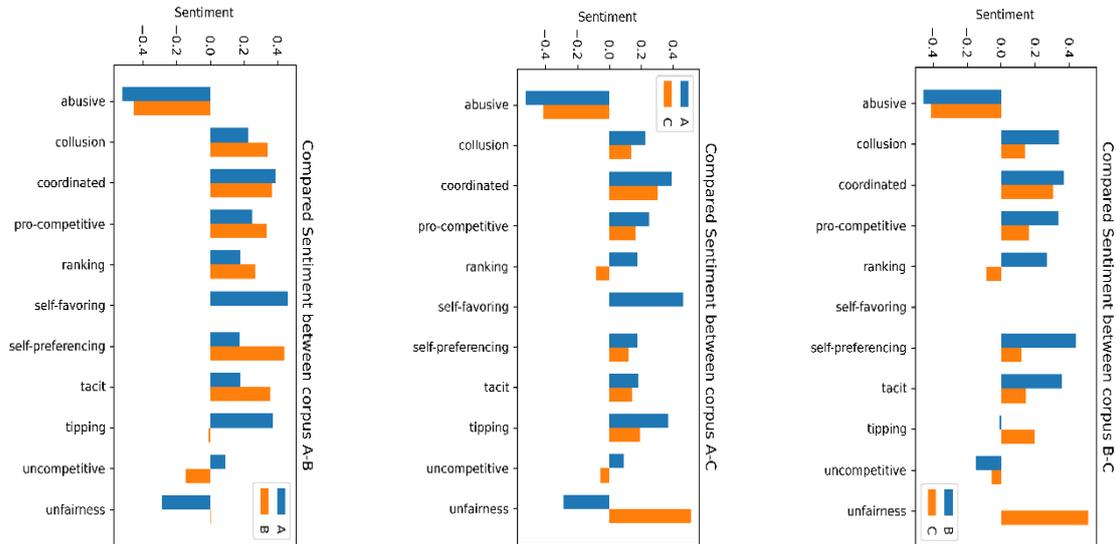


Figure 5: Sentiment Analysis. Compared Corpora Related to Anti-competitive Conduct



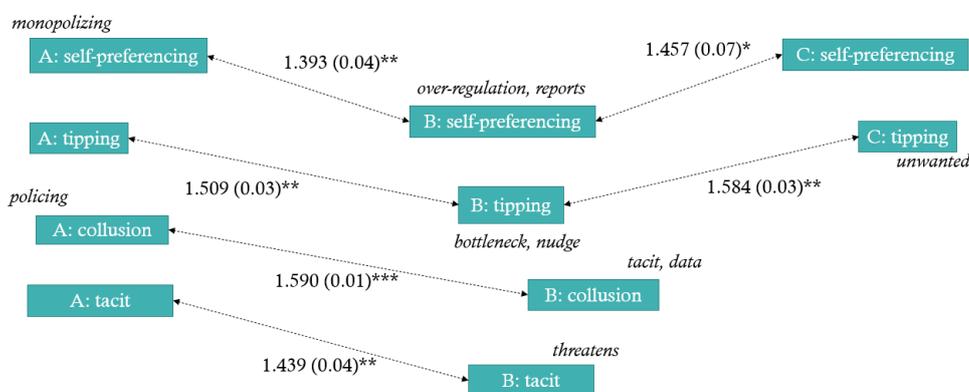
Moving on to terms related to potentially anti-competitive conduct (Fig. 4), it is remarkable to see that general terms like “uncompetitive” and “pro-competitive” are not used in the same way. The same holds true for the more drastic, but also commonly used term “abusive.”

Another interesting finding is that “pro-competitive” is closely associated with “user-friendly” for small companies and organizations (B), which could be interpreted as a hint that this group of stakeholders, in particular, perceive a strong link between consumer protection and friendliness (or fairness, in the DMA and DSA jargon) and competition. Individuals and micro-organizations (A), on the other hand, connect “pro-competitive” with the term “comparison”. This could emanate from the centrality of comparing offers to boost competition and

innovation, and seems to confirm the importance of provisions like article of the 6(1)d DMA⁷¹ or article 29 DSA.⁷²

Strictly related to that, is the statistically significant distance we find between the vectors of “self-preferencing” for corpuses A and B as well as for B and C. For corpus pair AC, the alternative term “self-favoring” is used differently as well, suggesting that it is the general concept behind these two terms that seems to be still elusive. In corpus A, “self-preferencing” is most closely related to “monopolizing,” (see Fig. 6) indicating that the anti-competitive outcomes of self-preferencing are a key concern for individuals and micro-organizations.

Figure 6: Distances and closest words for terms related to “anti-competitive practices”



Small companies, on the other hand, associate this practice with “over-regulation” and “reports.” This is interesting not only with a view to the prohibition on self-preferencing in Article 6(1)d of the DMA, but also in light of the comprehensive reporting duties on rankings in the DSA (Articles 12 and 29). Our results could be interpreted as a clue that small companies and organizations might fear comprehensive transparency and reporting duties, thus highlighting the need for a differentiated approach.⁷³

Another anti-competitive practice which yields a significant distance between corpuses A and B is “collusion.” For small businesses and organizations (B), this keyword is closely related to “tacit” which well-reflects the many debates on “tacit collusion,” driven both by the doctrine⁷⁴ as well as competition authorities,⁷⁵ and lawmakers.⁷⁶ Note that the term “tacit” itself is used in a similarly idiosyncratic manner, which might reflect the debate on what exactly constitutes an agreement or otherwise unlawfully coordinated behavior. The term “coordinated” is also used

⁷¹ DMA art. 6 §1(d) deals with self-preferencing in the context of ranking services (a practice addressed in the *Google Shopping* case, *supra* note 36).

⁷² *Supra* note 13.

⁷³ However, such a proportional approach is not easy to find, especially if there is no consensus on what “disproportionally” burdensome provisions look like. This seems to be the case at least for corpus pair AB.

⁷⁴ See e.g. Ariel Ezrachi & Maurice E. Stucke, *Sustainable and Unchallenged Algorithmic Tacit Collusion*, 17 NORTHWESTERN J. TECH. & INTELL. PROP. 217 (2020).

⁷⁵ See e.g. Bundeskartellamt & Autorité de la Concurrence (*supra* note 35).

⁷⁶ Crémer report, *supra* note 14, at 68.

differently by different stakeholders, although only small companies seem to connect it with tacit collusion. Interestingly, another term that shows up in the vicinity of “collusion” is “data,” which is in line with the conclusion drawn by many experts regarding the central role of data availability as an enabling factor of tacit collusion.⁷⁷

It is also worth noting that no reference is made in the DMA to “data” or “algorithmic collusion,” even though both locutions were referred to in the consultations. Given that other close relations exist with the words “harms” and “barred,” we could deduce a rather negative attitude towards the emergence of collusion in corpus B. Generally, collusion is relevant with a view to existing competition law (esp. Art. 101 TFEU), but was not picked up in the DMA, despite the long debate.

A potential avenue for anti-competitive conduct of online platforms that is being addressed by both the DSA (Article 29) and the DMA (Article 6(d)) are “rankings,” for which we find a statistically significant distance between corpuses A and C. A similar discrepancy was found for “tipping,” which is an important concept related to the need for ex-ante rules. In the small companies/organizations corpus, “tipping” is surrounded by the terms “bottleneck” and “nudge.” This mirrors the debate around gatekeepers, which are often considered to gain their dominant or quasi-dominant role from their position as bottlenecks, nudging consumers into certain choices.⁷⁸ Interestingly, our sentiment analysis (Fig. 5) shows that “tipping” is viewed more favorably by medium and big companies/organizations C (0.338) than by small companies/organizations B (-0.296). The negative attitude of small companies and organizations towards “tipping” might reflect the potentially even higher costs of uncompetitive markets small businesses face, because they might often depend on uncontested platform markets.

⁷⁷ See, for instance, Crémer report, *Ibid.*, at 8.

⁷⁸ There has even been empirical evidence for problematic nudges, so called “dark patterns.” Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty & Arvind Narayanan, *Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites*, 3 PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION, (CSCW) 81. Concerning bottlenecks, the UK Furman report incorporated the concept in its ‘significant market status’ threshold, which is comparable to the “gatekeeper” designation in the DMA. HER MAJESTY’S TREASURY, UNLOCKING DIGITAL COMPETITION, REPORT OF THE DIGITAL COMPETITION EXPERT PANEL (2019), <https://www.gov.uk/government/publications/unlocking-digital-competition-report-of-the-digital-competition-expert-panel> (accessed 25 February 2021), 10.

Regulatory Reform and Remedies

Figure 7: Projection of Aligned Word Embedding. Aligned Vector Space Model Corpora

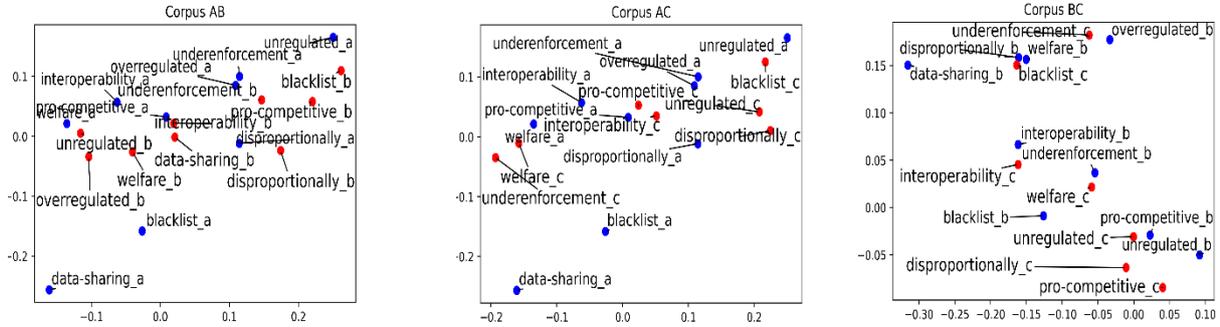
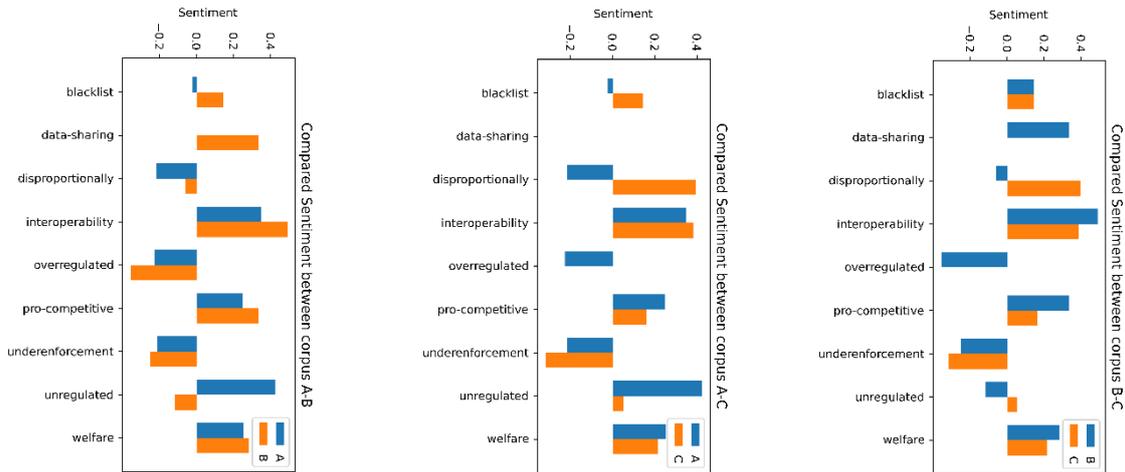


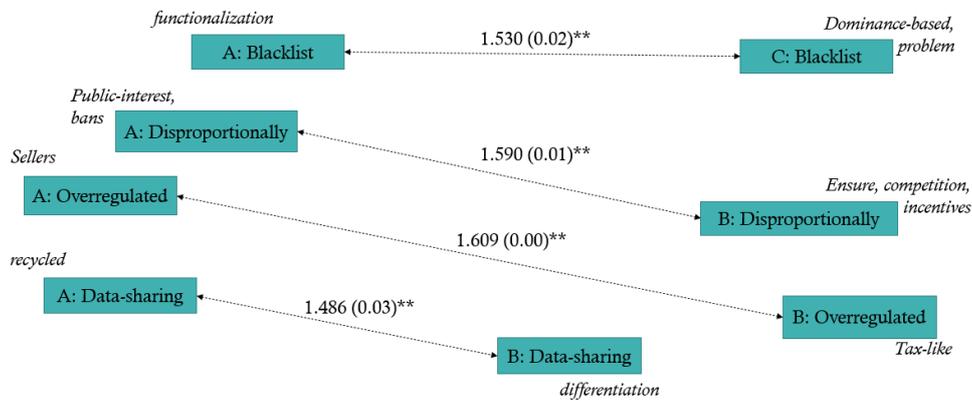
Figure 8: Sentiment Analysis. Compared Corpora Related to Regulatory Reform and Remedies



Having discussed concepts related to competition challenges on digital markets, the logical next question is: What to do about it? Regarding the general level of regulatory intensity, there seems to be no perfect alignment on the meaning of “overregulated,” with small organizations appearing to fear “tax-like” measures. At the same time, at least micro-organizations and individuals look to perceive a lack of regulation as “not sufficient.” Generally, for the term “unregulated” we find significant differences between all corpora. In a similar vein, we find a considerable distance between the vectors of “underenforcement” for Corpora A and B as well as A and C.

Looking at how stakeholder groups associate words (Fig. 9), while micro-organizations and individuals (A) seem to emphasize the importance of “unbureaucratic” procedures, medium and big organizations/companies (C) interestingly speak of “misconceptions” and “improvements.” While the precise meaning of these associations remains obscure, it is interesting to note that not only individuals and micro-entities but also bigger organizations seem to demand some kind of reform.

Figure 9. Distances and closest words for terms related to “regulatory reform” and “remedies”



Regarding the concrete measures of reform proposed by the Commission, we find that stakeholders have different understandings of three key concepts (Fig. 7). First, the term “blacklist,” which is highly relevant for the DMA,⁷⁹ is not used in the same way by micro entities as by medium/big organizations. Interestingly, the former mentioned the concept far less than the latter (1,972 occurrences in Corpus A, compared to 36,161 in Corpus C). This disparity could be explained by a certain sense of alarm on the side of larger entities that is hinted by the associated term “problem.” While this interpretation is highly tentative and would need to be confirmed by further analysis, the second closely associated term, “dominance-based” is a bit more telling and highly noteworthy. In fact, it might reflect the demand to set the threshold of application for new, stricter rules at the “dominance” threshold.⁸⁰ Linking this finding with our results for the terms “gatekeeper” (the approach chosen by the Commission) and “dominant,” it seems reasonable to conclude that the disagreement on the applicability *ratione personae* of new rules is also reflected in the use of certain pivotal terms related thereto.

Furthermore, our sentiment analysis (Fig. 8 above) shows that “blacklists” are seen more favorably by small, medium, and large companies than by individuals/micro entities. This might explain some of significant differences we find in the use of the term.

Second, moving from the “don’ts” of the blacklist part of the DMA to the positive obligations complementing these prohibitions, we find two noteworthy differences between Corpuses A and B. For instance, “interoperability” was used in a substantially different way, which might be a reflection of the debate on how far-reaching interoperability obligations should be.⁸¹ Interestingly, the focus on

⁷⁹ To be specific, DMA Article 5(1)(a), (d)-(f) and DMA Article 6(1)(a), (d)-(e) “blacklist” certain actions.

⁸⁰ See e.g., CERRE, DMA report, *supra* note 28, at 52.

⁸¹ Currently, Article 6(f) of the DMA only covers ancillary services, which many would like to see changed. Samuel Stolton, *EU SMEs in Bid for Greater Interoperability in Digital Markets Act*, EURACTIV, <https://www.euractiv.com/section/digital/news/eu-smes-in-bid-for-greater-interoperability-in-digital-markets-act/> (accessed 13/05/2021).

“reliability” and trustworthiness in the closely associated words in Corpus A matches the spirit of Article 6(i)(c) of the DMA, which grants gatekeepers the opportunity to restrict interoperability to the extent necessary to ensure “the integrity of the hardware or operating system provided.”

Lastly, we found significant differences in the use of the term “data-sharing.” While individuals and micro-organizations (A) emphasize the possibility to “recycle” the data gatekeepers are obligated to share under, e.g., Article 5(i)(g) and Article 6(i)(g) of the DMA, small companies and organizations (B) heed the need for “differentiation.” Potentially, this association is an expression of a fear of being overburdened with costly, potentially disproportional data-sharing duties. However, what exactly “disproportionally” burdensome remedies look like seems to be a matter of debate for itself, given that the term shows a significant difference between corpuses A and B.

III. NAMING IS TAMING? DRAWING LEGAL LESSONS FROM COMPUTATIONAL ANALYSES

Using computational tools to analyze the DSA and DMA consultation feedback documents allowed us to reach three results. First, we found that considering small and medium entities as a homogenous group of stakeholders might not always be recommendable, at least not for natural language data. In our computational analysis, it turned out that medium organizations are more comparable to big contributors than small ones. This highlights the need for data-driven procedures when determining the optimal units of analysis. As outlined below, there are several computational tools the Commission could use to respond to this need.

Second, our algorithmic analysis has shown that there are statistically significant differences between stakeholder groups in the use and understanding for some key concepts of competition policy. On the one hand, our analysis reproduced some of the results the Commission outlined in the summaries of the consultations. For instance, the differences in the terms “gatekeepers” or certain remedial strategies between different groups can be linked back to the debate on these concepts in the sense that they were not entirely uncontroversial.⁸²

On the other, we spotted substantial differences between stakeholders for terms the Commission perceived as uncontroversial. For example, the Commission concluded that “[t]he large majority of stakeholders believed that the proposed list of problematic practices, or “blacklist,” should be targeted to clearly unfair and harmful practices of gatekeeper platforms”; while we found that the use and understanding of “blacklist” differs significantly among different feedback contributor groups. The same holds for some important anticompetitive practices, such as “self-preferencing.”

The second finding of our computational analysis is that the consensus the Commission identified over the ex-ante measures proposed in the DSA and DMA

⁸² DMA, Explanatory Memorandum, 7-8.

might not be as unanimous as it seems at first sight: although stakeholders might say the same, they could mean different things.

These findings are relevant for two reasons. First, exposing “hidden misunderstandings” can enhance the quality of EU consultation processes. Since differences in understanding could lead to (undesirable) differences in implementation of those provisions that require stakeholders to act, it should be in the interest of the Commission to identify such differences before drawing conclusions from the consultations. Second, the findings of our analysis shine a new light on the scholarly debate about central terms of the DSA and DMA. While theoretical discussions are without a doubt the pivotal starting point for any kind of reform, our analysis adds some empirical insights regarding the clarity of certain concepts, which might in turn inspire new theoretical arguments.

Yet, our analysis certainly also has some limitations. First, from a technical point of view, our corpus is rather small and heterogenous due to the great number of different feedback contributors. One way to help mitigate this shortcoming would be to enlarge the corpus, on the one hand, and applying techniques like bootstrapping,⁸³ on the other. Second, we had to do some manual coding to select the most interesting terms for which we found significant distances. This part of the research process should be automatized or made more easily replicable in the future. Lastly, and most importantly: this analysis identified differences in understanding and use of some relevant terms, but cannot explain *why* they occur. However, although it does not allow us to empirically prove a causal link between how clearly defined a term is, either by law or jurisprudence, and inconsistencies in its use, our results do suggest that such a connection exists.

One idea could be to run the same algorithms on new data, namely the submissions to the current consultations.⁸⁴ Now that the term ‘gatekeeper’ is defined using at least some quantitative criteria, it would be interesting to see whether stakeholders use the term more consistently than before the proposal was published, which is the consultation phase we looked at. If we found that ‘gatekeepers’ is used more consistently, while e.g., self-preferencing, which was not clearly defined is still used inconsistently, it would be a very strong sign that such quantitative or at least clear criteria help to avoid misunderstandings.

⁸³ Trist'n Joseph, Bootstrapping Statistics. What it is and why it's used. (<https://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307>) Jun 17, 2020 (accessed August 10, 2021).

⁸⁴ We refer to the public consultation on the Data Act opened by the EC on June 3, 2021 and currently undergoing (https://ec.europa.eu/eusurvey/runner/Data_Act#, accessed on August 10, 2021) (covering many subjects and terms of the DSA and DMA, and addressing the same stakeholders. E.g. Section VI of the questionnaire explicitly refers to ‘gatekeepers’).

IV. CONCLUDING REMARKS

This paper set out to explore whether different stakeholders share a similar understanding of the many new competition challenges coming with the increased importance of digital markets. To do so, we contrasted the debate spurred by the consultation on the three EC inception impact assessments with the corresponding norms in the DSA and DMA proposals. We then employed computational tools to gain a fine-grained understanding of the stakeholders' feedback documents.

Analyzing replies to the EU Commission's public consultation, we find significant differences in stakeholders' use of central terms of competition law like for instance "gatekeepers," "procompetitive," "collusion," and "self-preferencing."

While we believe that discerning latent differences in the use of certain terms competition law is a crucial capability that could significantly enhance the consultation process, both lawmakers and legal scholars could benefit even further from quantitative text analysis if the tool we present in this paper is complemented by other NLP techniques.

Hence, the tools presented here should be seen as only a first building block of a fully-fledged NLP toolbox. For instance, "topic modeling"⁸⁵ could be used to get an intuitive understanding of which topics are the most relevant to stakeholders. Another powerful tool to be utilized more systematically in the future is "sentiment analysis". While we employed this technique to investigate the general attitude of each group of contributors towards certain words of interest, we did so by using a pre-trained model. While this analysis produced some interesting results, one could use the same idea to cluster statements based on the sentiment of a group towards a certain concept or proposal to get a better understanding of how supporters and critics of a proposal are distributed and what their main concerns and arguments are.⁸⁶

If statements that are inputs to regulation are clustered based on "document similarity measures",⁸⁷ this could help to perceive certain similarities or alliances between stakeholders, even across different groups like e.g., small companies and medium/large companies. As our analysis has shown, these clusters might not always look like what one would expect *prima facie*.

On a more legal ground, computational tools could be used to trace back the influence of certain stakeholders by identifying those statements which are the most similar to the rules the Commission decided to propose. This could allow to

⁸⁵ See David M. Blei, Andrew Y. Ng & Michael I. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

⁸⁶ See e.g., Shi Feng, Daling Wang, Ge Yu, Chao Yang & Nan Yang (2009). Sentiment Clustering: A Novel Method to Explore in the Blogosphere. In Q. Li, L. Feng, J. Pei, S. X. Wang, X. Zhou, & Q.-M. Zhu (Eds.), *Advances in Data and Web Management*. Springer, 332–344.

⁸⁷ See, e.g., Bambang Krismono Triwijoyo, & Kartarina Kartarina, (2019). Analysis of Document Clustering based on Cosine Similarity and K-Main Algorithms. *Journal of Information Systems and Informatics*, 1(2), 164–177. Gerhard Dannemann (2019). Comparative Law: Study of Similarities or Differences?, in Oxford Handbook of Comparative Law (2d ed.), M. Reimann and R. Zimmermann (eds.). Oxford: Oxford University Press, 390–422.

gain a precise understanding of why rules were drafted in a certain way and greatly help the interpretation of norms in light of their *telos* and their drafting history.

Consequently, computational analysis could uncover novel insights into the provenance of a provision: which stakeholders asked for it, where does it come from? Given that the analysis of the drafting history and objectives of a norm is an essential part of its exegesis, these insights are essential for legal scholars. And especially when it comes to proposals as complex as the DMA and DSA, they could be of great value both for the Commissions and legal scholars, and hence something to look forward to.

That might suggest both to further the inquiry, and to make an effort to spread a common understanding of relevant terms.

Appendix

Annex 1 – Results of preliminary data analysis to motivate corpus construction

Our clustering choice is based on two considerations: First, a qualitative analysis of the questionnaires accompanying the feedback documents⁸⁸ allowed us to get an understanding of which aggregation would cluster comparable feedback contributors together. Second, we conducted a quantitative analysis of the same questionnaires to ensure that our clusterization choices are solid. In particular, we sought to ensure that there is no statistically significant difference between *medium* and *large* entities in our sample since at least medium *companies* are often grouped with small, rather than large *companies*.

However, it needs to be noted that our feedback contributors are not only businesses, but also other types of organizations. This diversity could “smooth” the differences we would have expected to find if our sample included companies only. In fact, our qualitative analysis of the questionnaires suggested that medium entities in our sample are more comparable to large businesses/organizations.⁸⁹ To test the robustness of this perception, we analyzed the answers provided by medium and large entities to specific multiple choices questions.⁹⁰ We applied a Kolmogorov-Smirnov two-sample test⁹¹ to understand if there is a statistically significant discrepancy between the distribution of the answers of the two groups. If that was the case, we would assume that these answers must be considered as provided by two different populations, not allowing us to treat them as a unique cluster. The results of the test are shown in Figure 1:

⁸⁸ For an archive of the feedback materials, see European Commission, *Digital Services Act – Deepening the Internal Market and Clarifying Responsibilities For Digital Services*, Public Consultation 11 January 2021, <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12418-Digital-Services-Act-package-ex-ante-regulatory-instrument-of-very-large-online-platforms-acting-as-gatekeepers/public-consultation> (accessed 28 January 2021).

⁸⁹ While this could be due to the idiosyncrasy of our sample, this finding also corresponds with scholarly literature. See e.g. Ron Kemp & Clemens Lutz, *Perceived Barriers to Entry: Are There Any Differences Between Small, Medium-sized and Large Companies*, 3 INTERNATIONAL JOURNAL OF ENTREPRENEURSHIP AND SMALL BUSINESS 538 (2006).

⁹⁰ We selected said questions based on what could be considered interesting for our research. A full list of the questions we selected can be found in the appendix.

⁹¹ Joseph Lawson Hodges Jr, *The Significance Probability of the Smirnov Two-sample Test*, 3 MATEMATICA 469 (1958).

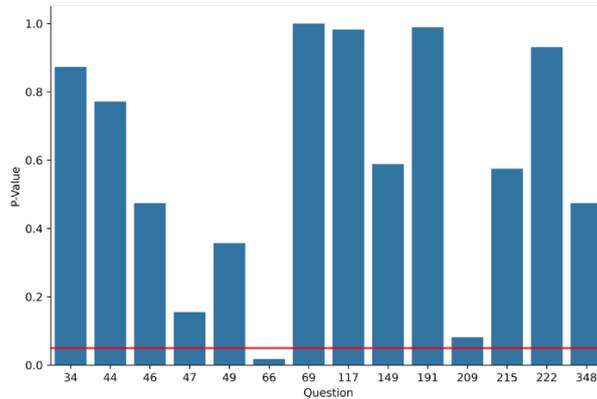


Figure 1: *p*-value distribution associated with the KS-two sample test applied to the answer distributions of the considered questions. Red line highlights our significant tolerance value of $p=0.05$

Even using a very high tolerance p -value level of 0.05, only question no. 66⁹² showed a statistically significant variation. This question alone however is mostly unrelated to our core research interest, and hence unlikely to compromise the validity of our clustering.

Annex 2 – A statistical test to identify semantic differences

We can model the relative distance d_t^{AB} of a word t in the corpus A and B be as:

$$d_t^{AB} = \gamma_t^{AB} + \mu_t^{AB} + u_t^{AB}$$

This takes into account a semantical term γ_t^{AB} , a non-semantical term (originated from the simple different words disposition in the two corpus) and a random term u_t^{AB} .

To isolate the semantic difference in the distances between words we found, we need to set two assumptions. Our first assumption is that words in the control vocabulary used for the Vector Space Alignment Transformation do not have a semantic difference, i.e., $\gamma_t^{AB} = 0$. This means we assume that stakeholders mean the same when they use words like “and” or “one.” Consequently, based on the distances between these “control” words, we can construct an empirical distribution of the non-semantic distance between words. In this manner, we can get an idea of what a distance would look like if there was no semantic difference.

In a next step, we can hence compare the distance we observe for certain words with what we would expect if there was no semantic difference. Our p -value represents the probability to have a distance equal or greater than what we observe if our null-hypothesis, that there is no semantic difference between the same term in different corpuses, is actually true. If this probability is small enough, we can reject this null hypothesis with a small possibility of error. This is to say that the

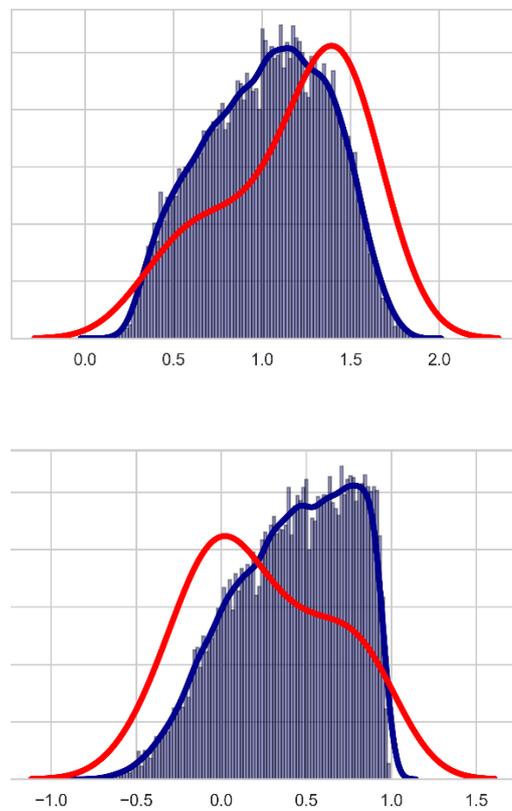
⁹² See Annex I.

particular word has, indeed, a statistically significant semantic difference in the two corpuses. A general acceptance value for the p-value is 0.05, which we will use as the critical threshold for our analysis.

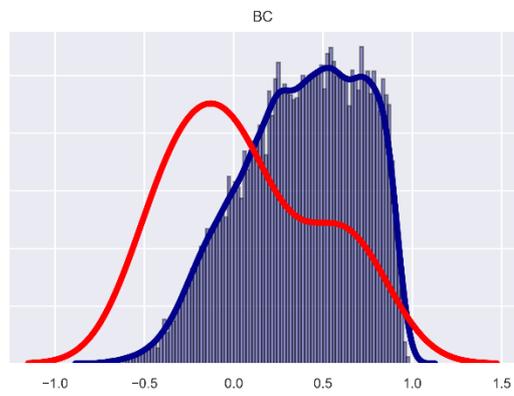
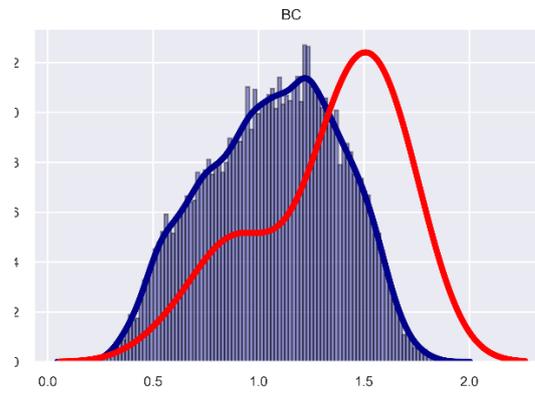
Annex 3 – Cumulative distribution of control distances

Figures 1 to 3 show the Cumulative Distribution of distances of control dictionary words (in blue) against cumulative distribution of distances of analyzed words (in red) for each corpus pair (i.e. corpus X against corpus Y).

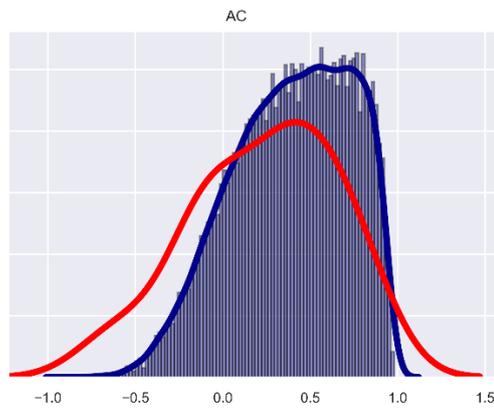
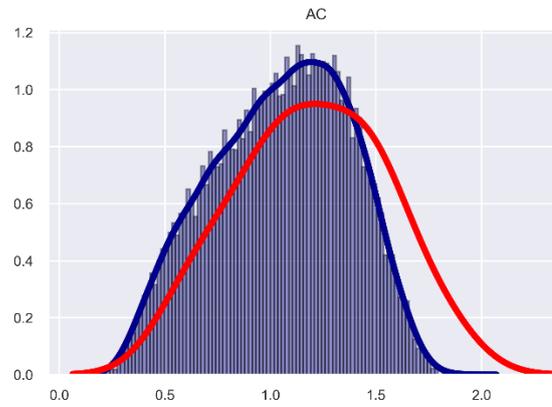
3.1 Corpuses AB - Cumulative distribution of control distances



3.2 *Corpus BC - Cumulative distribution of control distances*



3.3 Corpus Pair AC - Cumulative distribution of control distances



Annex 4 – Questions manually selected from the DSA questionnaire⁹³ to run the Kolmogorov-Smirnov two-sample test

Question 34: Did you ever come across illegal content online (for example illegal incitement to violence, hatred or discrimination on any protected grounds such as race, ethnicity, gender or sexual orientation; child sexual abuse material; terrorist propaganda; defamation; content infringing intellectual property rights, consumer law infringements)?

Question 44: Do you consider these measures⁽⁹⁴⁾ appropriate?

Question 46: If your content or offering of goods and services was ever removed or blocked from an online platform, were you informed by the platform?

Question 47: Were you able to follow-up on the information?

Question 49: If you provided a notice to a digital service asking for the removal or disabling of access to such content or offering of goods or services, were you informed about the follow-up to the request?

Question 66: Does your organisation access any data or information from online platforms?

Question 69: Do you use WHOIS information about the registration of domain names and related information?

Question 117: What information would be, in your view, necessary and sufficient for users and third parties to send to an online platform in order to notify an illegal activity (sales of illegal goods, offering of services or sharing illegal content) conducted by a user of the service?

- a) Precise location: e.g., URL
- b) Precise reason why the activity is considered illegal
- c) Description of the activity
- d) Identity of the person or organisation sending the notification. Please explain under what conditions such information is necessary
- e) Other, please specify

Question 149: In your view, is there a need for enhanced data sharing between online platforms and authorities, within the boundaries set by the General Data Protection Regulation? Please select the appropriate situations, in your view:

Question 191: Do you believe that in order to address any negative societal and economic effects of the gatekeeper role that large online platform companies exercise over whole platform ecosystems, there is a need to consider dedicated regulatory rules?

Question 209: Which, if any, of the following characteristics are relevant when considering the requirements for a potential regulatory authority overseeing the large online platform companies with the gatekeeper role:

- a) Institutional cooperation with other authorities addressing related sectors – e.g., competition authorities, data protection authorities, financial services authorities, consumer protection authorities, cyber security, etc.
- b) Pan-EU scope
- c) Swift and effective cross-border cooperation and assistance across Member States

⁹³ Note that the questions have been re-enumerated consecutively. The content remains the same.

⁹⁴ Question 43: 'What actions do online platforms take to minimize risks for consumers to be exposed to scams and other unfair practices (e.g. misleading advertising, exhortation to purchase made to children)?'.

- d) Capacity building within Member States
- e) High level of technical capabilities including data processing, auditing capacities
- f) Cooperation with extra-EU jurisdictions
- g) Other

Question 215: Taking into consideration the parallel consultation on a proposal for a New Competition Tool focusing on addressing structural competition problems that prevent markets from functioning properly and tilt the level playing field in favor of only a few market players. Please rate the suitability of each option below to address market issues arising in online platforms ecosystems. Please rate the policy options below from 1 (not effective) to 5 (most effective):

- a) Current competition rules are enough to address issues raised in digital markets
- b) There is a need for an additional regulatory framework Imposing obligations and prohibitions that are generally applicable to all large online platforms with gatekeeper power
- c) There is a need for an additional regulatory framework allowing for the possibility to impose tailored remedies on individual large online platforms with gatekeeper power, on a case-by-case basis
- d) There is a need for a New Competition Tool allowing to address structural risks and lack of competition in (digital) markets on a case-by-case basis.
- e) There is a need for combination of two or more of the options 2 to 4.

Question 222: When you see an online ad, is it clear to you who has placed the advertisement online?

Question 348: In your view, is there a need to ensure similar supervision of digital services established outside of the EU that provide their services to EU users?