Studies in Knowledge and Innovation
Vol. 2 No. 4A (2026)

Tied Timestamps Leave Consecutive Short-View Record Counts Undetermined

Cameron Richardson[1] · Trevor Gardner[2] · Benjamin Lewis[3]

[1] Switzerland · [2] Ireland · [3] New Zealand

Published: April 30, 2026 | Article 843

Abstract

An ordered-looking interaction table can leave most neighboring positions unresolved. We examine 225,406 eligible KuaiRand records from one fixed source date, retaining tied timestamps rather than breaking them arbitrarily. Exact optimization over the declared source-record orderings gives 28,329–65,727 consecutive short–short pairs among 224,521 adjacent slots. The corresponding share ranges from 0.126175 to 0.292743. A reference preserving each user-date block's label counts lies inside this interval, so the direction of excess adjacency is undetermined. Two convenient tie orders differ by only 97 pairs, offering little indication of that wider range. Timestamp semantics may concern requests rather than individual exposures. These are deterministic record-ordering diagnostics, without identification of physical viewing sequences, swipe events or playback-quality causes.

Tied Timestamps Leave Consecutive Short-View Record Counts Undetermined

Abstract

An ordered-looking interaction table can leave most neighboring positions unresolved. We examine 225,406 eligible KuaiRand records from one fixed source date, retaining tied timestamps rather than breaking them arbitrarily. Exact optimization over the declared source-record orderings gives 28,329-65,727 consecutive short-short pairs among 224,521 adjacent slots. The corresponding share ranges from 0.126175 to 0.292743. A reference preserving each user-date block's label counts lies inside this interval, so the direction of excess adjacency is undetermined. Two convenient tie orders differ by only 97 pairs, offering little indication of that wider range. Timestamp semantics may concern requests rather than individual exposures. These are deterministic record-ordering diagnostics, without identification of physical viewing sequences, swipe events or playback-quality causes.

1. A timestamp column does not supply every neighboring relation

Sequential analysis starts with an apparently simple operation: sort the records and inspect what follows what. When several rows share a timestamp, that operation introduces a decision. The resulting neighbors can reflect file order or another convenient tie break rather than information supplied by the source. A statistic of consecutive short records can then acquire a point value even though the recorded ordering constraints permit many values.

Our question is how much the joint share of consecutive below-three-second records can vary when equal-time groups are reordered. We hold the eligible records, their binary labels and the ordering of distinct timestamps fixed. We compute the exact minimum and maximum of the statistic across this declared class, including contributions between groups. The object is a source-record permutation problem. It does not presume that every row represents an independent exposure or that every permissible permutation is a possible physical viewing trajectory.

This boundary matters for interpreting rapid-viewing signals. A working-paper abstract separates quality-associated rapid exits from preference feedback using playback, network and device information with adjudicated evaluation (Yuan et al., 2026). Our public source contains neither those exit measurements nor those quality judgments. Its useful role is to show why unavailable sequence information cannot be replaced by an arbitrary record sort. The results do not validate that paper's private predictor or attribute short records to playback failure.

The distinction is part of a broader measurement problem. Numeric scales specify permissible operations but do not supply every substantive interpretation (Stevens, 1946). Measurement analysis connects an observable to the construct it is intended to represent (Jacobs and Wallach, 2021). Data-quality work also asks whether information is useful for its intended consumer, rather than only whether individual entries appear accurate (Wang and Strong, 1996). Here no record is declared wrong. The issue is the strength of a sequence claim that the available keys can support.

2. What the official source establishes and leaves open

KuaiRand combines standard recommendation records with random-item interventions and multiple feedback signals (Gao et al., 2022). We reuse the official, already acquired 1K release. Its sampled-user logs provide a more suitable recorded sequence than the Pure version, which removes interactions outside a restricted item pool. Nevertheless, a sampled-user recommendation archive is not automatically a complete account of physical viewing or off-platform activity.

The two intervention-period CSVs contain 6,657,061 standard and 43,028 random-source rows. Their actual extracted bytes and SHA256 digests were recorded, and the existing acquisition record retains the archive's verified provider checksum. Both files have nineteen columns, including time_ms, source date, tab, recorded viewtime and clip duration. We use only these fields, reindexed grouping keys and source-row locators. User profiles, captions, categories and original videos do not enter the analysis.

Official documentation describes time_ms as an interaction timestamp, without specifying a physical playback start or end (KuaiRand project, n.d.). The official repository's issue 4 explicitly asks whether it refers to a request or an exposure; the captured public thread contains no visible maintainer clarification. That unresolved question cannot be converted into an assertion that it is a request timestamp. It does mean that same-time rows may represent a recommended list or request, and their interpretation as independently timed exposures is unestablished.

The owner's reply in issue 3 gives an aggregate guide to interface scenarios: the most frequent tab corresponds to the single-column scenario. Tab 1 is most frequent in the two actual files, with 4,473,935 rows. We therefore analyze the recorded tab 1 subgroup. This combines an observed frequency with the author's guidance; it does not establish row-level interface ground truth, actual scrolling, a session boundary or a playback end.

The source's Zenodo metadata identifies CC BY 4.0, while its repository notice states CC BY-SA 4.0. Both notices and their provenance are retained with conservative attribution and share-alike handling. We do not invent a single resolved license or substitute a software license for the data terms. These source conditions are separate from the mathematical question addressed below.

3. The analyzed subsequence and its missing order

The primary date is the first captured source date, 20220422. It was fixed before any empirical joint-adjacency bound was computed. An earlier unexecuted candidate mentioned the full period; the assigned scope restricted the analysis to this inspected first date before results. Source and marginal-label support had already been examined, so this is a secondary analysis rather than a preregistered or outcome-blinded study.

The date contains 369,464 source rows from 919 reindexed user-date blocks. Eligibility requires recorded tab 1, finite nonnegative viewtime and clip duration strictly greater than 3,000 milliseconds. S denotes viewtime strictly below 3,000 milliseconds; N denotes other eligible records. Exactly 3,000 milliseconds belongs to N. We do not compare alternative thresholds, reweight users or infer a quality label from either state.

Filtering yields 225,406 records in 885 nonempty blocks: 83,600 S and 141,806 N. The S count includes 2,978 recorded zeros and 80,622 positive durations below the threshold. A zero is retained as a source value, without a claim that a video played and was immediately stopped. Seventeen records lie exactly at the threshold. Missing source values are not filled with zeros, and repeated selected keys are not silently merged into one presumed event.

The analyzed neighbors are positions in each **eligible-record subsequence**. Intervening ineligible records have been removed by that explicit definition. Consequently, adjacency here is not immediate adjacency in the entire original log, much less consecutive physical videos. The 919 observed blocks include 34 with no eligible record and twenty with one. Empty blocks have zero transition counts by definition, while both empty and single-record blocks have undefined transition rates because they contain no adjacent slot.

There are 865 multi-record blocks and 224,521 total adjacent slots, calculated as the sum of n-1 over nonempty blocks. Eligible records occupy 44,642 timestamp groups. Of these, 41,018 contain ties, comprising 221,782 records. Another useful structural count is 29,699 groups containing both S and N, involving 822 users. No unknown within-group order is recovered by applying a stable sorting routine.

Ordering scholarship distinguishes a declared relation from a convenient total ordering. Lamport's analysis makes ordering rules explicit in distributed events (Lamport, 1978); we do not infer a happens-before relation from these logs. Allen's interval framework also requires temporal relations that a bare interaction timestamp cannot supply (Allen, 1983). Process-mining methods depend on how event logs represent activities and traces (van der Aalst, 2016). Here the class is deliberately limited to permutations of source records under a chosen key constraint.

4. Coupling group endpoints gives exact record bounds

Within a user-date block, distinct timestamp groups retain increasing order. Every permutation of the binary records inside each group is allowed. This is a finite ordering class, related to the combinatorial idea of extending given order constraints (Stanley, 2011). Its boundaries are mathematical: unrecorded restrictions might narrow it, and a physical interpretation might fail altogether. We use it to quantify dependence on a record-sorting convention.

The count of S→S neighbors includes pairs inside an equal-time group and pairs crossing to the next group. Independent optimization inside each group is therefore insufficient. Choosing an arrangement that ends in S can change the best arrangement of the following group. Our dynamic program retains the last binary state of each processed prefix, using each current group's feasible first and last states and its internal count bounds.

For a group containing m short records, its internal S→S count equals m minus the number of S runs. The feasible number of runs depends on both endpoint states and the number of N records. At a group boundary, the recurrence adds one precisely when the previous endpoint and current first state are both S. This retains the information needed for the next step, following the dynamic-programming principle (Bellman, 2010). The same calculation is applied separately to the other three binary transitions.

We validated the primary implementation against exhaustive small examples, including 114 single-group endpoint cases and 819 coupled chains of one to three groups. Those examples are artificial mathematical checks, not empirical observations or simulated user behavior. They establish implementation agreement on the checked cases; real-data aggregation and saved results are checked separately. No downloaded repository program or learned model is run.

Figure 1 shows the relationship between the recorded groups and the reported effect. Proposition P1 concerns the width of the exact S→S share range. Proposition P2 asks whether that range determines the direction of excess against a declared fixed-marginal reference. Neither proposition concerns a physical swipe rate.

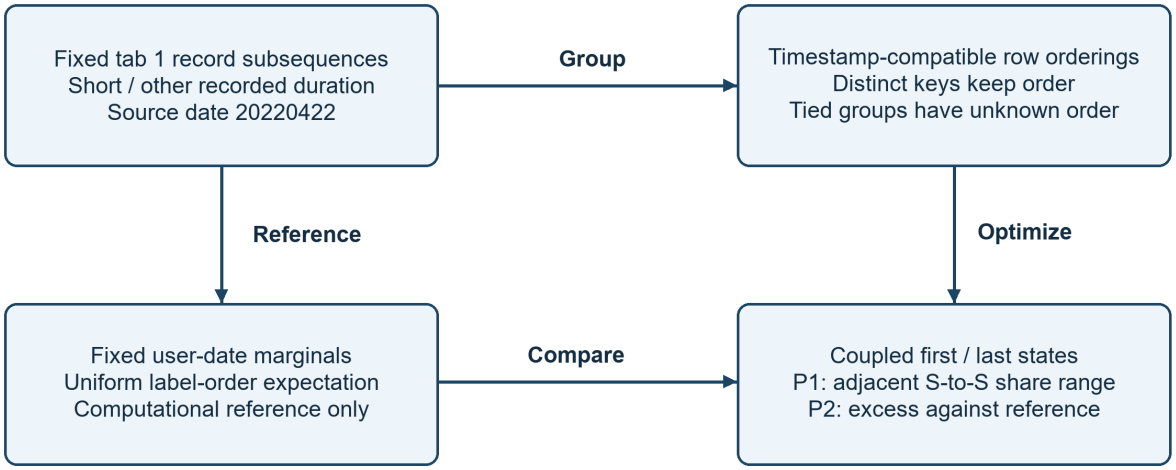


Figure 1. Research model. *Eligibility defines a recorded subsequence; timestamp groups define a mathematical ordering constraint. Coupled endpoint optimization produces finite ranges. The reference preserves each user-date block's short and nonshort counts. No arrow represents a recovered physical viewing order or a causal playback effect.*

5. Wide ranges coexist with similar convenience orders

The exact S→S count ranges from 28,329 to 65,727 over the declared source-record class. Dividing by the same 224,521 slots gives 0.126175 to 0.292743. The width is 0.166568, or 16.6568 percentage points. At the block level, 783 user-date blocks have more than one attainable S→S count. Thus the ambiguity is not confined to one unusually long record group.

Table 1. Separately optimized adjacent-record counts and shares.

Adjacent states	Minimum count	Maximum count	Minimum share	Maximum share
S → S	28329	65727	0.126175	0.292743
S → N	17474	54966	0.077828	0.244815
N → S	17516	55090	0.078015	0.245367
N → N	86138	123803	0.383652	0.551409

S is recorded viewtime below three seconds; N is another eligible record. Every share uses 224,521 total eligible-subsequence slots, not the number of S predecessors. Each row is optimized separately with coupled group endpoints. The four minimum entries, or the four maximum entries, do not generally form one jointly attainable transition matrix. Ranges are exact deterministic ordering ranges for 225,406 records in 885 nonempty user-date blocks, not confidence intervals.

For comparison, Table 2 uses source-file/row order within ties and then reverses that order inside each group. Both are permitted convenience orderings. Neither is privileged as the observed physical sequence. Each complete matrix sums to the fixed total of adjacent slots, providing an audit of state and boundary accounting.

Table 2. Two legal convenience orders, neither treated as truth.

Convenience ordering	N → N	N → S	S → N	S → S	All slots	S → S share
Source-row tie order	100112	41171	41059	42179	224521	0.187862
Reversed source-row tie order	100026	41240	41173	42082	224521	0.187430

The two S→S counts differ by 97, a change of only about 0.0432 percentage points. Their closeness could seem reassuring if they were the only ordering checks. Yet their agreement leaves almost the entire optimization range unexplored. Trying two convenient sorts is a different question from finding the full range under the stated constraints.

This distinction resembles the broader partial-identification lesson that conclusions depend on the maintained assumptions (Manski, 2003; Tamer, 2010). We borrow that discipline without claiming an identified set for real-world viewing. Our target statistic is defined on hypothetical source-row permutations. Confidence procedures for partially identified population parameters address an additional sampling problem (Imbens and Manski, 2004); no such confidence procedure is performed here.

6. A fixed-marginal reference does not resolve the sign

For a nonempty block with n eligible records and m short records, uniform permutation of its labels has expected S→S count m(m-1)/n. When n is at least two, there are n-1 adjacent slots, and an ordered pair of distinct positions is short-short with probability m(m-1)/(n(n-1)). Empty and single-record blocks have no adjacent pair and contribute zero expected count. This analytic reference preserves each block's marginals rather than pooling all users into one label distribution.

Unlike the optimization class, the reference does not preserve the short/nonshort composition of each timestamp group. It is a benchmark for adjacency, not the mean over the permitted within-group orderings. Keeping that distinction explicit prevents a convenient reference distribution from being mistaken for the source's observation mechanism.

Summing those expectations gives 38,788.974699 pairs, or a reference share of 0.172763. Fractional expected counts are valid expectations, not fabricated source rows. The reference falls inside the exact S→S range. Subtracting it gives an excess-share interval of **-0.046588 to +0.119980**. Therefore, even the sign of this record-ordering excess is not fixed by the maintained constraints.

Permutation arguments have a long history in order-sensitive statistics (Wald and Wolfowitz, 1940). Practical permutation methods condition on specified structures (Good, 2005), while randomization-test interpretation requires attention to the mechanism justifying a reference distribution (Edgington and Onghena, 2007). Our calculation supplies no claim that actual viewing labels were randomly assigned, no experimental treatment and no p-value. It is a transparent comparator to the deterministic extrema.

Leaving out each observed user-date block changes the lower excess endpoint between -0.047074 and -0.046058, and the upper endpoint between 0.119599 and 0.120338. All these intervals still contain zero. These are composition checks on one captured date, not independent replications or a sampling interval. More importantly, they cannot settle the unresolved meaning of the timestamp field.

7. What the record diagnostic can support

The unresolved order is not equivalent to an unobserved scalar outcome. Missing-data methods depend on the nature and mechanism of absent information (Little and Rubin, 2002); we neither impute an order nor assume its absence is random. Eligibility also selects a subset of the source records. Selection can change an analytical population (Heckman, 1979), and our thresholds do not justify claims about all users, dates or application events.

Even an interval wholly above the reference would describe this record statistic under its assumed ordering class. It would not identify a quality-driven behavioral mechanism. Causal inference requires a different relationship between observations and interventions (Holland, 1986). Likewise, burst models use substantive timing and temporal structure (Kleinberg, 2003); a count of neighbors in a filtered record list is not a measurement of physical rapid swiping.

Digital traces create opportunities for social research (Lazer et al., 2009), while their observation and design choices remain consequential (Salganik, 2017). Reviews of social-data biases describe how collection processes and intended questions can diverge (Olteanu et al., 2019). In this case, the appropriate response is to narrow the claim to the source-record object. The unanswered field question is not evidence that the dataset's authors supplied erroneous records.

Data-quality assessment should specify which required information is available (Pipino et al., 2002), and broader data-quality frameworks distinguish dimensions and purposes (Batini and Scannapieco, 2016). The relevant missing ingredient here is a supported ordering or event definition for the desired physical claim. Replacing it with a numerically tidy ordering would change the question rather than supply that ingredient.

8. Reproducibility and conclusion

Dataset documentation should retain the actual field meanings and intended-use limits (Gebru et al., 2021). Distinguishing source rows, timestamp groups and user-date transition counts makes the analytical units explicit (Wickham, 2014). The saved calculation follows the principle of inspectable computational results (Peng, 2011), with intermediate decisions retained for verification (Sandve et al., 2013). Source hashes, eligibility, block counts, all four transition bounds and the two complete diagnostic matrices accompany the result.

For this fixed first date, consecutive below-three-second record shares are weakly determined by the chosen timestamp constraints. Two nearly matching convenience orders conceal a much wider exact range, and the fixed-marginal reference lies inside it. That finding is useful as an ordering-sensitivity diagnostic. It does not reconstruct physical exposures, establish swipe events or determine why a recorded viewtime is short. Those claims require temporal and playback evidence absent from the present fields.

References

1. Batini, Carlo; Scannapieco, Monica. (2016). Data and Information Quality: Dimensions, Principles and Techniques. Data-Centric Systems and Applications. https://doi.org/10.1007/978-3-319-24106-7
2. Good, Phillip. (2005). Permutation, Parametric and Bootstrap Tests of Hypotheses. Springer Series in Statistics. https://doi.org/10.1007/b138696
3. Holland, Paul W.. (1986). Statistics and Causal Inference. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1986.10478354
4. Manski, Charles F.. (2003). Partial Identification of Probability Distributions. Springer Series in Statistics. https://doi.org/10.1007/b97478

5. Tamer, Elie. (2010). Partial Identification in Econometrics. Annual Review of Economics. https://doi.org/10.1146/annurev.economics.050708.143401

6. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

7. Heckman, James J.. (1979). Sample Selection Bias as a Specification Error. Econometrica. https://doi.org/10.2307/1912352

8. Sandve, Geir Kjetil; Nekrutenko, Anton; Taylor, James; Hovig, Eivind. (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1003285

9. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847

10. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

11. Yuan, Yuewei; Xu, Tianyi; Huang, Jialei; Yin, Jiayang. (2026). Playback Quality Anomalies Associated with Rapid Swiping in Short Videos: Model-attributed Association Estimation and Risk Prediction. SSRN working paper 7393398. https://doi.org/10.2139/ssrn.7393398.

12. Stanley, Richard P.. (2011). Enumerative Combinatorics. Cambridge University Press. https://doi.org/10.1017/cbo9781139058520

13. Wang, Richard Y.; Strong, Diane M.. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. https://doi.org/10.1080/07421222.1996.11518099

14. Gao, Chongming; Li, Shijun; Zhang, Yuan; Chen, Jiawei; Li, Biao; Lei, Wenqiang; Jiang, Peng; He, Xiangnan. (2022). KuaiRand: An Unbiased Sequential Recommendation Dataset with Randomly Exposed Videos. Proceedings of the 31st ACM International Conference on Information & Knowledge Management. https://doi.org/10.1145/3511808.3557624

15. Wald, A.; Wolfowitz, J.. (1940). On a Test Whether Two Samples are from the Same Population. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177731909

16. Olteanu, Alexandra; Castillo, Carlos; Diaz, Fernando; Kıcıman, Emre. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data. https://doi.org/10.3389/fdata.2019.00013

17. Lamport, Leslie. (1978). Time, clocks, and the ordering of events in a distributed system. Communications of the ACM. https://doi.org/10.1145/359545.359563

18. Imbens, Guido W.; Manski, Charles F.. (2004). Confidence Intervals for Partially Identified Parameters. Econometrica. https://doi.org/10.1111/j.1468-0262.2004.00555.x

19. Lazer, David; Pentland, Alex; Adamic, Lada; Aral, Sinan; Barabási, Albert-László; Brewer, Devon; Christakis, Nicholas; Contractor, Noshir; Fowler, James; Gutmann, Myron; Jebara, Tony; King, Gary; Macy, Michael; Roy, Deb; Van Alstyne, Marshall. (2009). Computational Social Science. Science. https://doi.org/10.1126/science.1167742

20. Edgington, Eugene; Onghena, Patrick. (2007). Randomization Tests. Chapman and Hall/CRC. https://doi.org/10.1201/9781420011814

21. Salganik, Matthew J.. (2017). Bit by Bit: Social Research in the Digital Age. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691158648/bit-by-bit

22. KuaiRand project. (n.d.). KuaiRand: dataset documentation, release license and field discussions. Official repository README, LICENSE and issues 3-4. Accessed 16 November 2026. https://github.com/chongminggao/KuaiRand

23. Pipino, Leo L.; Lee, Yang W.; Wang, Richard Y.. (2002). Data quality assessment. Communications of the ACM. https://doi.org/10.1145/505248.506010

24. van der Aalst, Wil. (2016). Process Mining. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

25. Allen, James F.. (1983). Maintaining knowledge about temporal intervals. Communications of the ACM. https://doi.org/10.1145/182.358434

26. Wickham, Hadley. (2014). Tidy Data. Journal of Statistical Software. https://doi.org/10.18637/jss.v059.i10

27. Bellman, Richard E.. (2010). Dynamic Programming. Princeton University Press. https://doi.org/10.1515/9781400835386

28. Jacobs, Abigail Z.; Wallach, Hanna. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445901

29. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677

30. Kleinberg, Jon. (2003). Bursty and Hierarchical Structure in Streams. Data Mining and Knowledge Discovery. https://doi.org/10.1023/a:1024940629314