Journal of Data Efficiency
Volume 1 · Number 1 · 2026

Neighborhood Breadth and Median Aggregation in a Beijing PM2.5 Withholding Study

Nathan Green[1] · Jacob Johnston[2] · Evan Anderson[3]

[1] United States · [2] United Kingdom · [3] Australia

Original publication: October 24, 2026 | Article 585

Abstract

More observed stations did not produce the most accurate consensus estimate in this retrospective exercise. We withheld each Beijing station in turn at 15,718 later-period hours with complete measurements across twelve sites. Three neighbors selected using earlier correlations yielded an equal-day, equal-site mean absolute error of 12.74 micrograms/m3, compared with 15.81 for all eleven available neighbors. Median aggregation reduced error in the broad neighborhood but increased the observed error in the three-neighbor set. A factorial regression with week-clustered uncertainty retained this interaction. These findings concern controlled single-station withholding and do not establish performance during real network outages.

Exploratory public-data research draft. Human scientific and submission review remain pending.

Neighborhood Breadth and Median Aggregation in a Beijing PM2.5 Withholding Study

Exploratory public-data research draft. Human scientific and submission review remain pending.

Abstract

More observed stations did not produce the most accurate consensus estimate in this retrospective exercise. We withheld each Beijing station in turn at 15,718 later-period hours with complete measurements across twelve sites. Three neighbors selected using earlier correlations yielded an equal-day, equal-site mean absolute error of 12.74 micrograms/m3, compared with 15.81 for all eleven available neighbors. Median aggregation reduced error in the broad neighborhood but increased the observed error in the three-neighbor set. A factorial regression with week-clustered uncertainty retained this interaction. These findings concern controlled single-station withholding and do not establish performance during real network outages.

1. A neighbor budget changes the information being combined

An air-quality network can observe a common pollution episode without every station representing the same local concentration. Pooling more stations may stabilize an estimate while moving it toward a citywide pattern that is less suitable for a particular site. Robust aggregation poses a related choice: reducing the influence of a remote response may help in a broad pool, but discarding the magnitude ordering among three particularly relevant responses can lose useful information. Neighborhood breadth and aggregation rule therefore deserve a joint evaluation.

This article asks how a three-neighbor budget changes reconstruction error relative to using every other station, and whether replacing a mean by a median changes that tradeoff. The observations are actual hourly PM2.5 measurements from the UCI Beijing Multi-Site Air Quality release (Chen, 2017). The associated assessment in Zhang et al., 2017 addresses broader questions about air-quality trends and meteorological adjustment. Our release contains twelve stations, and our outcome is reconstruction error. It should not be described as the broader source study's full monitoring network or as an assessment of pollution-policy effectiveness.

The comparison is intentionally conditional. At each evaluated hour, every station has an observed concentration. We hide one site's response from its reconstruction rule while retaining the actual value for evaluation. This creates a common benchmark on which four policies receive the same information availability apart from the chosen neighbor budget. It does not reproduce missingness caused by a power failure, severe weather, communication congestion, or a damaged monitor.

2. Network representation and the aggregation proposition

Graph-signal processing relates multivariate measurements to explicit network structure (Shuman et al., 2012; Ortega et al., 2017). We construct a directed neighborhood by ranking earlier pairwise concentration correlations. A neighbor is therefore a historically correlated site, not necessarily the geographically closest monitor. All eleven other sites form the broad comparison pool. The three highest-ranked sites form the restricted pool, with ties resolved by station name.

The target work, Peng et al., 2024, introduces a normalized sign-based adaptive method for robust graph-signal prediction, including a multivariate-temperature application. It makes robustness and efficient estimation relevant concerns for graph observations. Our question isolates two simpler

components that an application must still choose: the available neighbor set and the aggregation of its standardized responses. We execute neither the target's adaptive update nor its normalization procedure, so the present errors provide no replication or ranking of that algorithm.

P1 concerns the error difference between the restricted and broad neighborhoods under mean aggregation. Its sign is an empirical question because selecting correlated sites trades breadth against relevance. P2 concerns an interaction: the median-versus-mean difference can depend on neighborhood size. The robust-estimation principles in Huber, 1964 and Rousseeuw, 1984 motivate sensitivity to unusual observations, but a median is not guaranteed to minimize absolute reconstruction error after a particular scaling and graph selection.

The factorial structure appears in Figure 1. The two policy dimensions operate on a fixed set of withheld station-hours. Training-period location and scale statistics map every available neighbor onto a common standardized scale, followed by a map back into the focal station's concentration units. The evaluation compares complete deterministic policies rather than treating individual neighbor readings as independently randomized treatments.

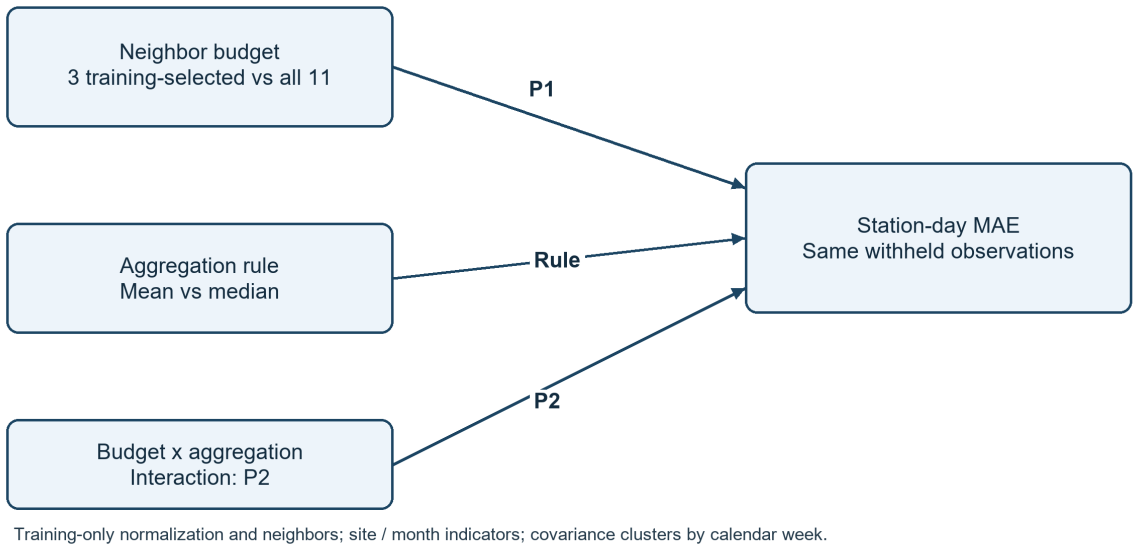


Figure 1. P1 compares three versus eleven neighbors for the mean rule. P2 is the neighborhood-by-median interaction. Training-only normalization and station/year-month indicators support a matched descriptive comparison; arrows are not causal pollution mechanisms.

3. What is actually inside the downloaded archive

The official outer ZIP includes a second archive named PRSA2017_Data_20130301-20170228.zip, a map image, and two ancillary CSV files. The ancillary files are not station pollution records; one contains stock-price columns. We use only the twelve station CSVs in the named inner archive. This file-level check matters because choosing the first CSV automatically would silently analyze a different subject while retaining the air-quality dataset's name.

Each station supplies 35,064 hourly timestamps, giving 420,768 original station-hour rows. We verify unique timestamps within station and align them on a common time index. The measured period runs from March 2013 through February 2017. Missing concentration values are retained as missing during preprocessing. Across all original station-hours, the PM2.5 missing fraction is approximately 2.08%; that marginal percentage should not be mistaken for the fraction of hours with all twelve sites available.

Records before 1 January 2015 form the training period, comprising 16,104 native timestamps. Means and population standard deviations use each station's available earlier values. Pairwise training correlations use available pairs with at least 1,000 observations. No later concentration determines a scaling coefficient or neighbor ranking. We check that every training scale is finite and positive before producing any reconstruction.

The remaining period has 18,960 native timestamps. Requiring all twelve actual concentrations leaves 15,718 hours spread across 773 days. This selection excludes hours with any site missing. The missing-data framework in Little and Rubin, 2002 explains why the resulting conditional benchmark need not describe an outage population. We make no claim that the retained complete hours are a random sample of all operational conditions.

4. Four reproducible reconstruction policies

For a focal station, we standardize every other site's contemporaneous concentration using that site's training mean and standard deviation. The restricted policies take the first three neighbors from the frozen correlation ranking; the broad policies use all eleven. Within each set, we calculate either the arithmetic mean or the median. We then multiply that standardized consensus by the focal site's training standard deviation and add its training mean.

No focal concentration enters its own reconstruction. No negative prediction is clipped after examining its error, and no validation outcome determines a policy-specific scaling. These choices make the comparison easy to reproduce, while leaving possible improvements outside the experiment. The measurement-error account in Carroll et al., 2006 is relevant because normalization does not turn observed concentrations into error-free latent values. It changes their scale, not their epistemic status.

The restricted policy bundles two decisions: using fewer sites and choosing those sites by their earlier correlations. It is not a randomized test of the number three against the number eleven while holding neighbor composition otherwise fixed. A random three-site pool, a geographic three-site pool, or a wider pool with fitted weights could perform differently. The estimated contrast belongs to the named composite policy. Describing it simply as evidence that less information is always better would erase the selection mechanism responsible for defining which information remains.

Standardization creates another interpretive boundary. We combine departures from each neighbor's earlier level in units of its earlier variability, then express the result on the focal site's historical scale. Thus the broad mean is not the raw citywide mean concentration. A site's typical level and spread are intentionally restored after aggregation. This can remove persistent level differences while leaving differing responses to particular episodes. The experiment assesses how the remaining normalized responses combine; it does not show that the raw concentrations are exchangeable across locations.

Pairwise training correlations are estimated on the observations available for each pair. Their eligible hours can differ because training readings are missing. The minimum-pair count provides a basic data requirement but does not eliminate uncertainty in the ranking or establish a stable graph across pollution regimes. That uncertainty is held fixed in the primary policy comparison. A new evaluation could re-estimate the graph on repeated earlier periods and then test subsequent periods, keeping all graph selection upstream of each evaluation window.

For each station-day-policy combination, we average absolute errors over the eligible hours on that day. All twelve stations and four policies use those same hourly positions. Each policy therefore supplies 9,276 station-day means, and the factorial regression contains 37,104 rows. Days with fewer eligible hours

receive the same day weight as more fully observed days. That is an explicit equal-day estimand rather than a pooled-hour accuracy score.

The forecasting discussion in Hyndman and Koehler, 2006 helps keep the named error measure and weighting visible. The scoring framework in Gneiting and Raftery, 2007 further distinguishes point-estimate evaluation from assessment of a predictive distribution. We calculate no probability interval for the missing station, no distributional calibration score, and no threshold-exceedance sensitivity. A policy with smaller mean absolute error need not dominate every operational loss.

5. A matched factorial regression

The response is station-day MAE in micrograms/m3. Regressors include a restricted-neighborhood indicator, a median indicator, and their interaction, plus station and year-month intercepts. The reference policy is the eleven-neighbor mean. Because every policy appears at every retained station-day, the main contrasts have a direct matched interpretation; the intercept is a reference station-month value rather than the overall benchmark mean.

We estimate the linear contrast model with ordinary least squares and covariance clustered by calendar week. There are 114 week clusters. All stations, policies, and days in the same week can share dependent disturbances. Clustered inference follows the dependence concerns in Cameron and Miller, 2015, while White, 1980 describes the separate issue of heteroskedasticity. Weather persistence across week boundaries and the finite selection of this network remain limitations of this uncertainty calculation.

The training-test split preserves time ordering, consistent with the validation arguments in Roberts et al., 2017. We did not select the budget by searching later errors across many values. Nevertheless, this is an exploratory analysis rather than a registered confirmatory test. If additional budgets are tried on this period, model-selection concerns such as those in Varma and Simon, 2006 apply to any best-performing choice subsequently reported.

The executed contrasts appear in Table 1. The interaction asks whether the median effect changes when the neighborhood is restricted. It should not be read as a causal effect of physically removing stations, because availability is controlled retrospectively and the actual measurements are held fixed.

Table 1. Station-day MAE regression with station and year-month indicators. Reference policy: eleven-neighbor arithmetic mean. Standard errors cluster by 114 calendar weeks, with a finite-sample adjustment and t(113) intervals. All policies use the same 15,718 complete test hours; N denotes policy-specific station-day rows.

Model	Term	Estimate (SE)	95% CI	N
Daily-site MAE policy contrast	Three minus eleven neighbors: P1	-3.0700 (0.2593)	[-3.5837, -2.5563]	37104
Daily-site MAE policy contrast	Median minus mean, eleven neighbors	-0.3571 (0.0418)	[-0.4400, -0.2743]	37104
Daily-site MAE policy contrast	Neighborhood x median: P2	0.6444 (0.0717)	[0.5023, 0.7865]	37104
Daily-site MAE policy contrast: fit	R-squared = 0.2144	-	-	37104

6. Breadth and robustness do not move together uniformly

The broad-neighborhood mean has MAE 15.8139, whereas the three-neighbor mean has MAE 12.7439. The estimated restricted-neighborhood contrast is -3.0700, with standard error 0.2593 and interval from -3.5837 to -2.5563. In this conditional benchmark, the smaller historically selected neighborhood is more accurate on the declared loss and weighting scale. This is an observed data-budget comparison, not evidence of a hardware or communication-speed improvement.

Using the median in the eleven-neighbor pool gives MAE 15.4568, a reduction of 0.3571 from its mean counterpart. The corresponding interval is -0.4400 to -0.2743. Within the restricted pool, median MAE is 13.0312, or 0.2873 above its mean counterpart. The interaction coefficient is 0.6444, with interval from 0.5023 to 0.7865. Thus the measured median-versus-mean contrast changes materially with the pool definition.

The model R-squared is 0.2144. Its role is to summarize variation in the saved station-day errors, not to report the fraction of pollution dynamics explained or the accuracy of a deployed sensor. Agreement concepts in Bland and Altman, 1986 caution against replacing an interchangeability question by a correlation. Likewise, the reliability guidance in Koo and Li, 2016 does not allow these deterministic policy replications to be counted as independent measurements of station reliability.

7. Relation to learned imputation and network assessment

Several established imputation approaches address richer problems than the present consensus calculation. Cini et al., 2021 develops graph-based multivariate time-series imputation, and Cao et al., 2018 uses a bidirectional recurrent approach. Yoon et al., 2018 and You et al., 2020 provide different generative or graph-representation routes to missing-data handling. These works motivate reporting the masking regime and available information carefully; their results are not imported into our comparison, and none is executed on these station files.

Tabular methods such as chained equations (Buuren and Groothuis-Oudshoorn, 2011) and MissForest (Stekhoven and Bühlmann, 2012) also show that reconstruction is a modeling choice. We did not compare those alternatives, tune a learned imputer, or use future values to fill a present gap. A wider method comparison would need a separate validation budget and common data-access rules. The simple policies here are useful precisely because the observed neighborhood and aggregation differences can be traced without attributing them to an opaque training procedure.

A model of natural outages would need information about why readings disappear. Causal-graph reasoning (Greenland et al., 1999) helps identify that missingness may share causes with the difficult conditions being predicted. Transport considerations in Stuart, 2010 similarly limit transfer from complete hours to an outage population. Even a large complete-case sample does not resolve that mismatch by itself.

8. Reporting the benchmark without turning it into a universal rule

An influence audit inspired by Cook, 1977 could investigate which days or station profiles account for unusually large errors. Such inspection should preserve the original comparison and label any subsequent exclusions as exploratory. If many neighborhoods, loss functions, and station-specific hypotheses are screened, multiplicity should be addressed using assumptions appropriate to their dependence. Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 illustrate why that choice cannot be reduced to adding significance symbols to a large table. The present intervals are unadjusted descriptive uncertainty summaries for the stated contrasts.

We retain the two nested archive identities, station timestamps, scaling rules, selected neighbor names, and station-day policy errors. Documentation in the spirit of Gebru et al., 2021 makes the exclusion of unrelated outer CSVs visible. The Wilkinson et al., 2016 principles support retaining identifiers and reusable metadata, while source attribution preserves the release's stated terms. These records make it possible to change a policy transparently without pretending that a new policy creates a new set of observed concentrations.

9. Conclusion

On the selected complete hours of this twelve-site Beijing release, three historically correlated neighbors outperform an eleven-site mean for equal-day, equal-site absolute reconstruction error. Median aggregation has a favorable measured contrast in the broad pool and an unfavorable observed contrast in the restricted pool. Neighborhood breadth and aggregation should therefore be evaluated jointly. Whether the same relationship survives real simultaneous outages, different station networks, or an adaptive sign-based estimator remains an open empirical question.

References

1. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
2. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.
3. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.
4. Yoon, Jinsung; Jordon, James; van der Schaar, Mihaela. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. arXiv:1806.02920. https://arxiv.org/abs/1806.02920.
5. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
6. Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
7. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
8. Chen, Song. (2017). Beijing Multi-Site Air Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5RK5G.
9. Greenland, Sander, Judea Pearl, and James M. Robins. "Causal Diagrams for Epidemiologic Research." *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.
10. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.
11. Cao, Wei; Wang, Dong; Li, Jian; Zhou, Hao; Li, Lei; Li, Yitan. (2018). BRITS: Bidirectional Recurrent Imputation for Time Series. arXiv:1805.10572. https://arxiv.org/abs/1805.10572.
12. Cini, Andrea; Marisca, Ivan; Alippi, Cesare. (2021). Filling the G_ap_s: Multivariate Time Series Imputation by Graph Neural Networks. arXiv:2108.00298. https://arxiv.org/abs/2108.00298.
13. Buuren, Stef van; Groothuis-Oudshoorn, Karin. (2011). mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software. https://doi.org/10.18637/jss.v045.i03.

14. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

15. Ortega, Antonio; Frossard, Pascal; Kovačević, Jelena; Moura, José M. F.; Vandergheynst, Pierre. (2017). Graph Signal Processing: Overview, Challenges and Applications. arXiv:1712.00468. https://arxiv.org/abs/1712.00468.

16. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

17. You, Jiaxuan; Ma, Xiaobai; Ding, Daisy Yi; Kochenderfer, Mykel; Leskovec, Jure. (2020). Handling Missing Data with Graph Representation Learning. arXiv:2010.16418. https://arxiv.org/abs/2010.16418.

18. Stekhoven, Daniel J.; Bühlmann, Peter. (2012). MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr597.

19. Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

20. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

21. Zhang, Shuyi; Guo, Bin; Dong, Anlan; He, Jing; Xu, Ziping; Chen, Song Xi. (2017). Cautionary tales on air-quality improvement in Beijing. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. https://doi.org/10.1098/rspa.2017.0457.

22. Benjamini, Yoav and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.

23. Peng, Changran; Yan, Yi; Kuruoglu, Ercan E.. (2024). Adaptive Graph Normalized Sign Algorithm. arXiv:2405.04107. https://arxiv.org/abs/2405.04107.

24. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

25. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

26. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

27. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

28. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.

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

30. Shuman, David I; Narang, Sunil K.; Frossard, Pascal; Ortega, Antonio; Vandergheynst, Pierre. (2012). The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. arXiv:1211.0053. https://arxiv.org/abs/1211.0053.