Testing compatibility of causal models with (historical) data under missingness not at random

Nick Gardner

Stanford University

November 14, 2022

Motivations

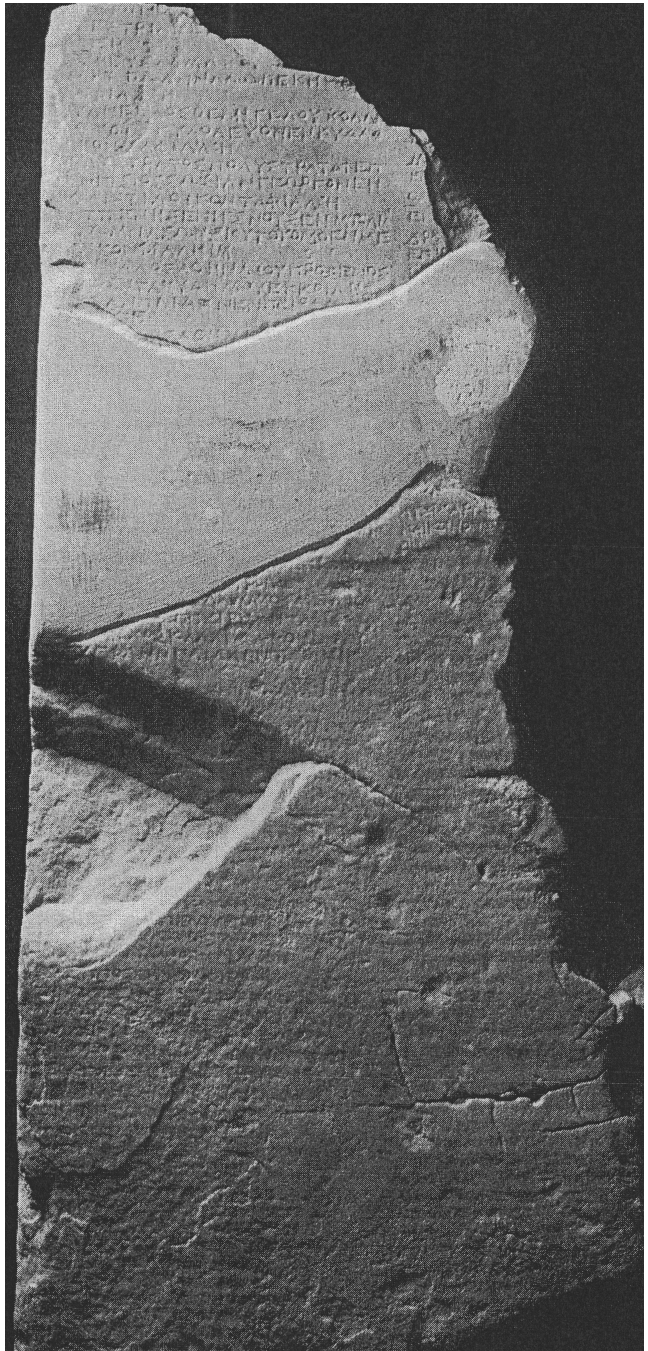
- (theory) Testability of causal DAGs is a fascinating set of problems that are still not well understood except in the case where all variables are fully observed
- (application) Causal modeling of missing data is great for research that works with historical data

Historical data: variables often (very) partially observed

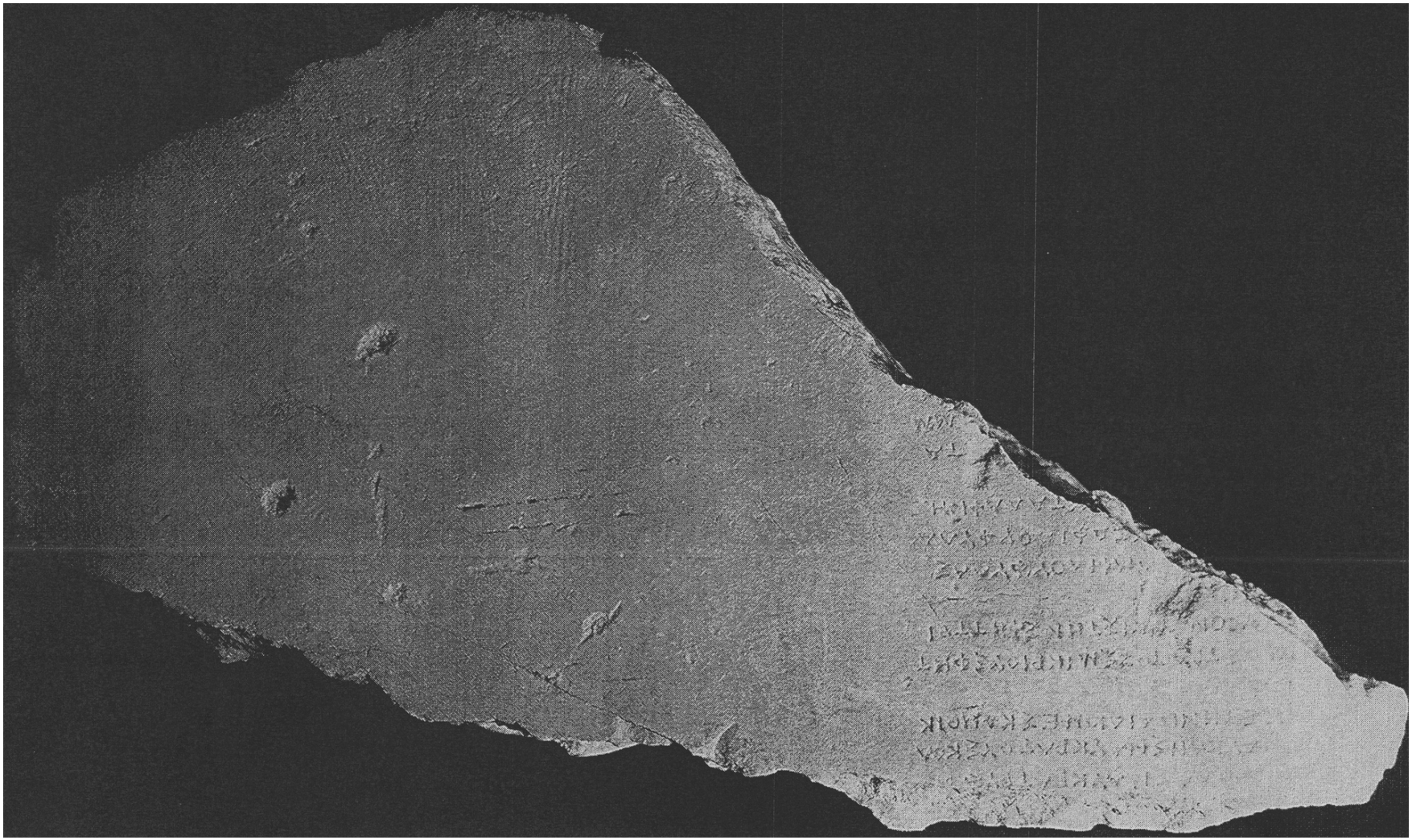
```
[– – – – – – – – – – – ἀποφυγ]-
.....13.....ος Σφ-
[ήττιον φιά σ]ταθμ Η· νν
[...8.... τα]λασι ἐν Κ-
[υδαθ οἰκοῦ ἀπ]οφυγοῦ
[....10.... Εὐ]θυκλέ-
[ου..6... φι]άλ σταθ: Η·
....9.... σιδηρο ἐν
[..6... οἰκ]ῶ ἀποφυγὼ
....9.... Λυσανίου
[...8.... φ]ιάλ σταθ :Η·
...8.... κ[ο]λλεψ Ἀλω-
[πεκ οἰκῶ] ἀποφυγὼ νvv
...8....ένη Ἀριστο-
[.... Παλ?]λη, φιά σταθ :Η·
...7... νευρορά ἐν Σ-
[καμ οἰκ]ῶ ἀποφυγὼ νvv
..6...α Πολυρήτ[ο]υ ν
...7... φιά σταθ:Η· νν
col. II.18      – – – – – τι[ον] Εὐ[κρά]-
 [τ]ους Ἐπικη, φιάλ στ[α :Η]·
[––––––]
[Σ]ωτ[η]ρίδης ὀνηλάτ [ἐν]
 Διομεί οἰκῶν ἀποφυ[γ]
 [Ἀ]ντιμένην Πιστοκλ[έ]-
 ου Κηφισιέ, φιά στα[θ :Η]·
```

```
Νικήρατος Νικηράτου Μελιτ,
Φείδιππος Σωσιδήμου Ξυπε
Στρατονίκην ἐμ Μελ οἰκ ταλα,
φιά :Η·
––––––
Νικήρατος Νικηράτου Μελιτ,
Φείδιππος Σωσιδήμου Ξυπετ
Ἀριάνθην ἐμ Με οἰκ ταλα, φι :Η·
––––––
Λυσιάδης Χίωνος Ἀλωπεκ
Σωστράτην ταλασιουργ ἐμ Μ οἰκ,
φι :Η·
[Κ]αλλίας Καλλικράτους Ἀφιδ
..στον ἐγ Κολλυ οἰκ ὀνη, φι :Η·
– – κλῆς Ἀριστοφάνους Ἀχαρ
– – – ἐμ Μ [οἰκ ταλα]σιουρ, φι :Η
– – – – – – – – – – – ου Λευ
– – – – – – – – – – – φι :Η·
```

Historical data: variables often (very) partially observed

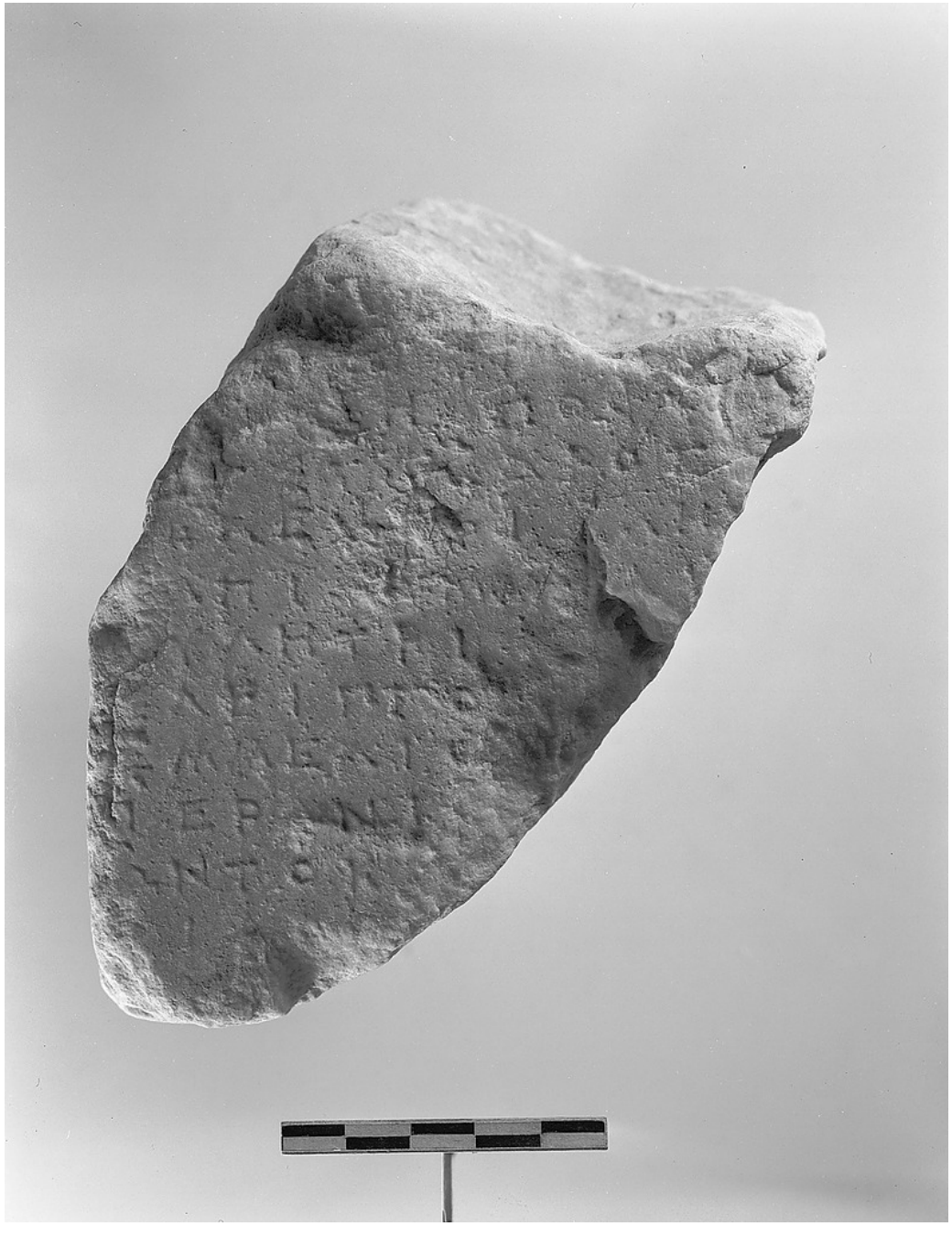


Historical data: variables often (very) partially observed



Historical data: variables often (very) partially observed

Historical data: variables often (very) partially observed



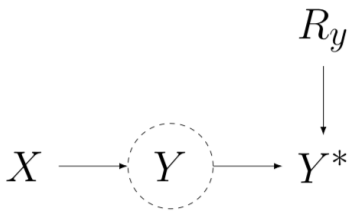
Today's Talk

Part 1: Overview of Missingness Graphs and MCAR/MAR/MNAR

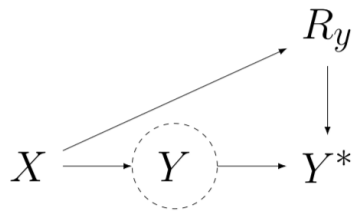
Part 2: Testability Under Missingness

Part 3: Modeling historical data with m-graphs

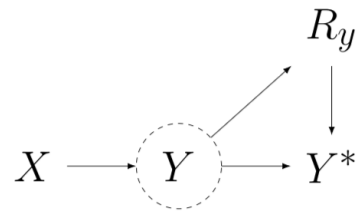
Missingness Graphs: Picture to Have in Mind



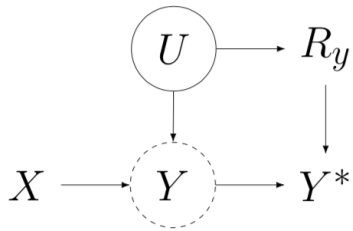
(a) MCAR



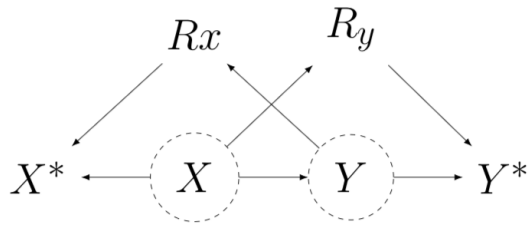
(b) MAR



(c) MNAR (self-censoring)



(d) MNAR (unobserved confound)



(e) MNAR (no self-censoring)

Missingness Graphs

Missingness graphs (m-graphs) are causal DAGs developed to model missing data problems [4, 3]. **Nodes** (variables) come in five different types:

- $\boldsymbol{V_o}$ – a set of fully observed variables
- $\boldsymbol{V_m}$ – a set of partially observed variables
- $\boldsymbol{U}$ – a set of completely unobserved variables (latent variables, hidden variables)
- $\boldsymbol{R}$ – a set of indicator variables representing missingness mechanisms that **cause** values in our data to be missing.
- $\boldsymbol{V^*}$ – a set of **proxy variables**, what we actually observe of the partially observed variables.

We refer to variables in $\boldsymbol{V_o} \cup \boldsymbol{V_m}$ as **substantive variables** (i.e. the variables we are primarily interested in modeling).

Missingness Graphs

One V_i^* is associated with every V_i in the set of partially observed variables $\boldsymbol{V_m}$. The value of each proxy variable is (non-stochastically) determined by the values of its associated missingness mechanism R_{V_i} and underlying variable V_i

$$v_i^* = f(r_{V_i}, v_i) = \begin{cases} v_i & \text{if } r_{V_i} = 0, \\ m & \text{if } r_{V_i} = 1. \end{cases}$$

The **edges** (causal influences) in an m-graph are usually subject to the following additional restriction: missingness indicators causally influence only proxy variables and other missingness indicators, not substantive or unobserved variables. Graphically, an edge from $\boldsymbol{R}$ can only go to $\boldsymbol{V}^*$ or stay inside $\boldsymbol{R}$ (without creating a cycle).

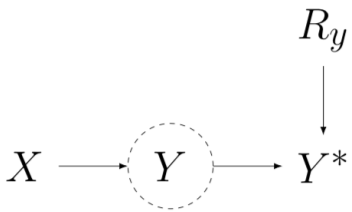
MCAR, MAR, and MNAR

MCAR: Data are "missing completely at random" or MCAR if $\boldsymbol{V_o} \cup \boldsymbol{V_m} \perp\!\!\!\perp \boldsymbol{R}$ holds in the m-graph.

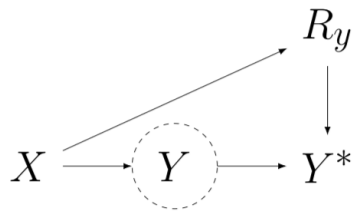
MAR: Data are "missing at random" or MAR if $\boldsymbol{V_m} \perp\!\!\!\perp \boldsymbol{R} | \boldsymbol{V_o}$ holds in the m-graph.

MNAR: Data that are not MCAR or MAR are "missing not at random" or MNAR. Some data are systematically missing in the following strong sense: there is no collection of fully observed variables which we can stratify on to render (within each stratum/level) the missingness mechanisms independent of the partially observed variables. The missingness is "not at random (conditionally or unconditionally)."

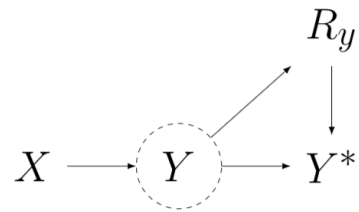
MCAR, MAR, MNAR in m-graphs



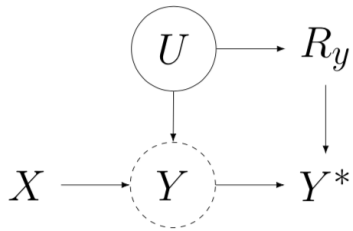
(a) MCAR



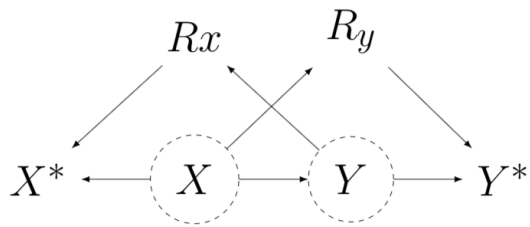
(b) MAR



(c) MNAR (self-censoring)



(d) MNAR (unobserved confound)



(e) MNAR (no self-censoring)

Recoverability and Estimation vs. Testability and Testing

- **Recoverability:** Assuming a causal model is correct, which of its parameters can in principle be estimated given unlimited sample data (which parameters have consistent estimators)?
- **Estimation:** What are efficient methods for estimating recoverable parameters from finite samples?
- **Testability:** Can the overall assumption that a given model is correct be tested, in principle given unlimited sample data? Can the specific qualitative assumptions that make up a model be tested individually (so that if a model is incorrect, we could identify and repair the assumptions that made it incorrect)?
- **Testing:** What are efficient methods for testing compatibility of models with finite samples of data?

Testability under missingness: open questions, future directions

- Complete algorithm for finding **conditional independence constraints** (i.e. *testable* conditional independence claims) entailed by an m-graph about the observed data distributions it models [2, 3]

Testability under missingness: open questions, future directions

- Complete algorithm for finding **conditional independence constraints** (i.e. *testable* conditional independence claims) entailed by an m-graph about the observed data distributions it models [2, 3]
- Complete algorithm for finding **Verma or other equality constraints**

Testability under missingness: open questions, future directions

- Complete algorithm for finding **conditional independence constraints** (i.e. *testable* conditional independence claims) entailed by an m-graph about the observed data distributions it models [2, 3]
- Complete algorithm for finding **Verma or other equality constraints** [9, 8, 7] in missingness graphs. See recent most recent work by Nabi and Bhattacharya on testability under missingngess [5]
- Any algorithm for finding **instrumental or other inequality constraints** ("Bell inequality style" constraints) in missingness graphs (moving beyond latent variable models where they were first discovered/studied [6, 1, 10]).

Testability under missingness: open questions, future directions

- Complete algorithm for finding **conditional independence constraints** (i.e. *testable* conditional independence claims) entailed by an m-graph about the observed data distributions it models [2, 3]
- Complete algorithm for finding **Verma or other equality constraints** [9, 8, 7] in missingness graphs. See recent most recent work by Nabi and Bhattacharya on testability under missingngess [5]
- Any algorithm for finding **instrumental or other inequality constraints** ("Bell inequality style" constraints) in missingness graphs (moving beyond latent variable models where they were first discovered/studied [6, 1, 10]).
- More **efficient tests** for the above constraints. Recent work explores tests based on weighted likelihood ratios and odds-ratio parameterizations of joint distributions [5]

Testability under missingness: open questions, future directions (cont'd)

- Algorithms for **causal structure learning under missingness**: given data with missing values, what is the set of all (or at least many of) the m-graphs compatible with the data? Is there a parsimonious/meaningful representation of this set that provides a unified characterization of the observed data? Structure learning under measurement error: [11]
- (speculative) Relate m-graph/causal inference work on recoverability, testability, and structure learning, to the **methods of information theory and coding theory** developed to study analogous problems: error correction and/or detection algorithms for variety of noise and erasure channels (classical, quantum, hybrid); algorithms for learning channels from observed input-output data.

MCAR, MAR, and MNAR Inscriptions

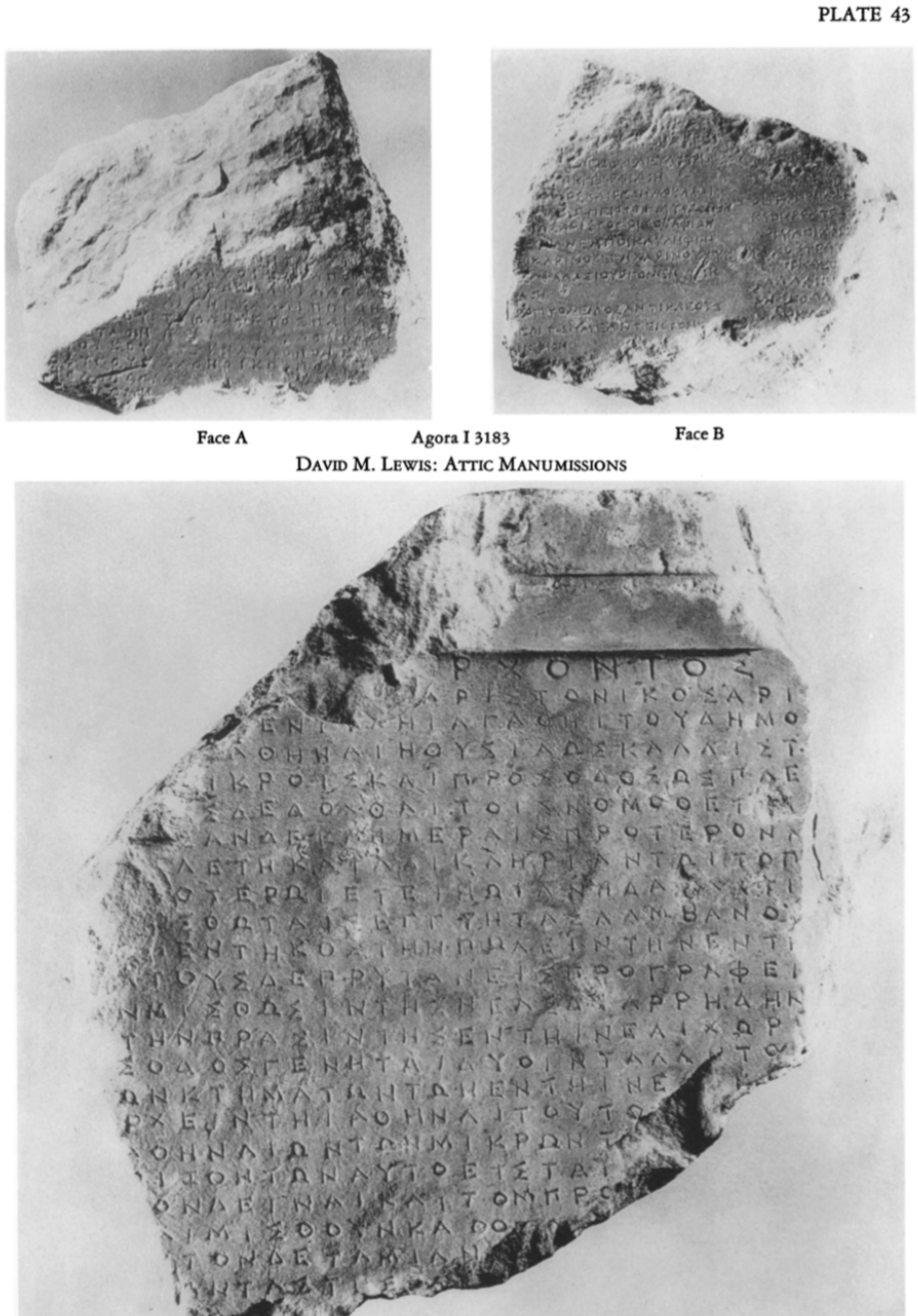
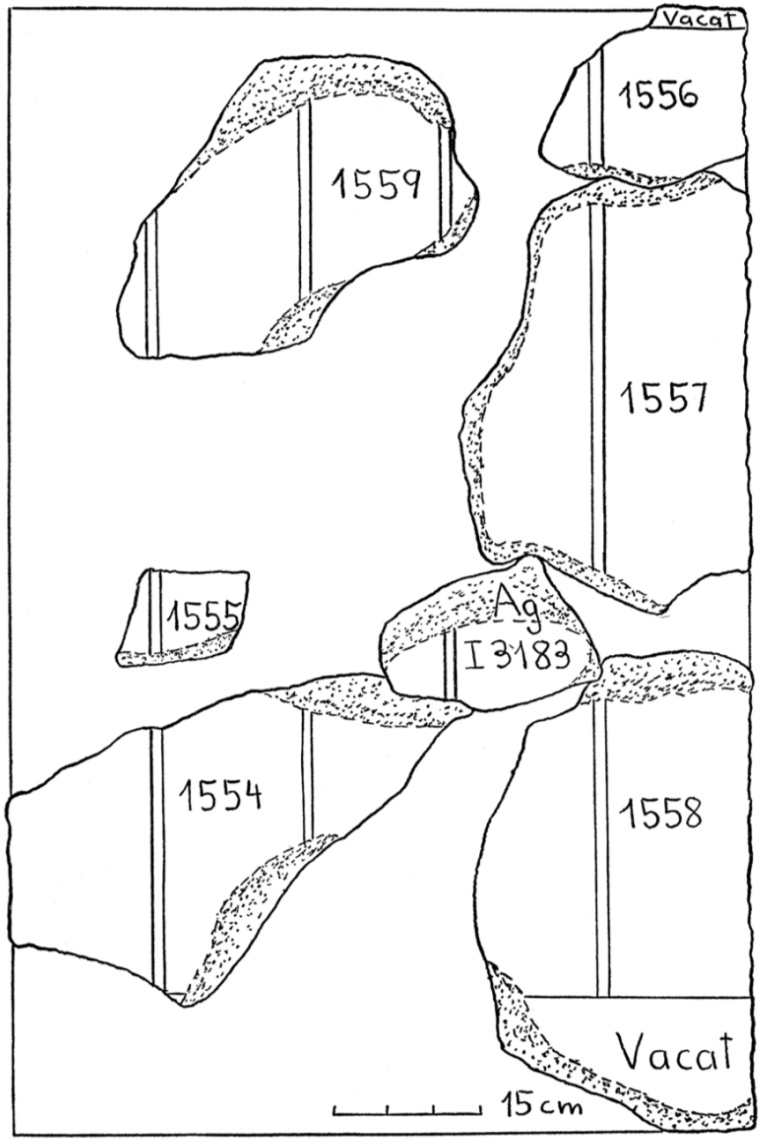


PLATE 43

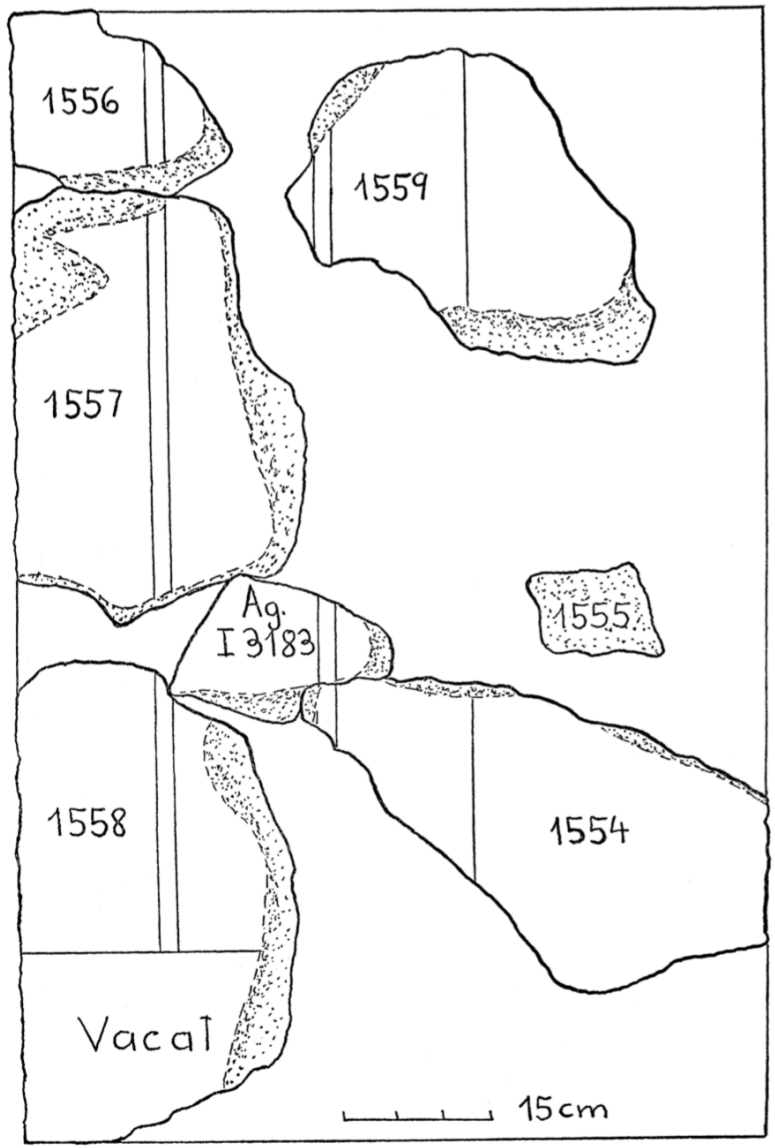
Face A Agora I 3183 Face B

DAVID M. LEWIS: ATTIC MANUMISSIONS

MCAR, MAR, and MNAR Inscriptions



FACE A

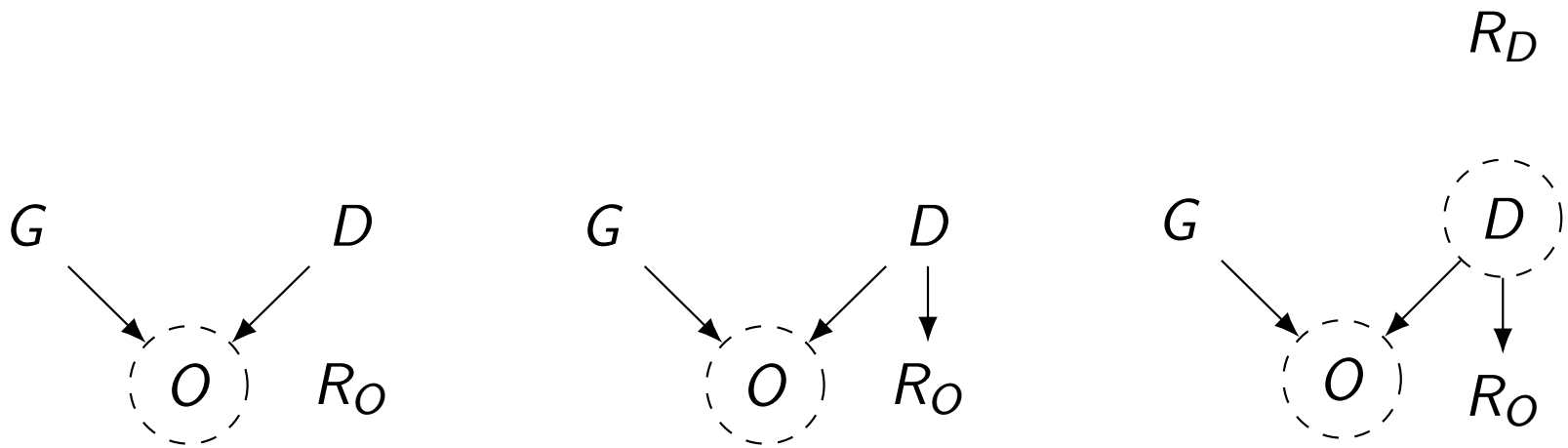


FACE B

MCAR, MAR, and MNAR Inscriptions

Modeled variables

- G: Gender
- O: Occupation
- D: District ("deme", one of 139 geographical subregions in Attica)



(a) MCAR (b) MAR (c) MNAR

Dissertation: Using m-graphs on real historical data

- What proportion of the total working population engages in which types of work?
- what is the distribution of culturally salient "types" of people (men/women, citi- zens/metics/slaves, children/young-adults/full-adults/elders) into types of work?
- what is the geographical distribution of types of work?

Thank you!

References

[1] Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 48–55, 2001.

[2] Karthika Mohan and Judea Pearl. On the testability of models with missing data. In *Artificial Intelligence and Statistics*, pages 643–650. PMLR, 2014.

[3] Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

[4] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. *Advances in neural information processing systems*, 26, 2013.

[5] Razieh Nabi and Rohit Bhattacharya. On testability and goodness of fit tests in missing data models. *arXiv preprint arXiv:2203.00132*, 2022.

[6] Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 435–443, 1995.

[7] Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

[8] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 519–527, 2002.

[9] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270, 1990.

[10] Linbo Wang, James M Robins, and Thomas S Richardson. On falsification of the binary instrumental variable model. *Biometrika*, 104(1):229–236, 2017.

[11] Yuqin Yang, AmirEmad Ghassami, Mohamed Nafea, Negar Kiyavash, Kun Zhang, and Ilya Shpitser. Causal discovery in linear latent variable models subject to measurement error. *arXiv preprint arXiv:2211.03984*, 2022.