**APPENDIX A**

COMPUTATIONAL APROACHES TO ARTISTIC SIMILARITY

As alluded to in our Data section, the assorted art sale data we collected from Blouin Artinfo was too large and messy to manually extract artwork information. In particular, the information on each piece’s materials was given in the form of somewhat unstructured text data, often with typographical errors. We suspect that Blouin provides several text fields for data entry labelled “Materials”, “Measurements”, and “Markings”, within which the specialists may write a freeform description. Below follows a sample of what this looks like after we collected it:

Materials: canvas Measurements: 23.62 in. (60.00 cm.) (height) by 28.74 in. (73.00 cm.) (width) Description: certificate from ComitA Picabia included Markings: signed and dated '1906'

Materials: pen and brown ink over counter-print of sanguine Measurements: 7.40 in. (18.80 cm.) (height) by 4.45 in. (11.30 cm.) (width)

Materials: black pencil heghtened with white chalk on blue paper Measurements: 9.45 in. (24.00 cm.) (height) by 4.92 in. (12.50 cm.) (width) Condition: foxing and folds

Materials: panel Measurements: 10.43 in. (26.50 cm.) (height) by 15.35 in. (39.00 cm.) (width) Markings: signed, l.l.

Materials: grey wash and brown wash on traces of black pencil, heightened with watercolor and white gouache Measurements: 9.25 in. (23.50 cm.) (height) by 6.50 in. (16.50 cm.) (width) Markings: old catalogue inscriptions verso 'no. 159' and 'no. 121'

Materials: black pencil, blending stump heightened with white chalk Measurements: 8.58 in. (21.80 cm.) (height) by 6.69 in. (17.00 cm.) (width) Description: oval

Materials: brown wash on paper Measurements: 5.12 in. (13.00 cm.) (height) by 5.91 in. (15.00 cm.) (width) Size Notes: sight Markings: signed and dated l.r. 'Deveria 1821'; stamped (joined) 'JR' l.r.

For this thesis, we only used simple keyword-based extraction. For example, to determine the presence for a signature, we would filter for observations with handcrafted keywords such as “signed,” “signature,” and so forth. This simple method was sufficient for our goal of detecting anchoring cross-effects in our assorted art dataset. However, our measures of similarity are far removed from the raw text, since after extracting attributes from keywords, we must fit a hedonic regression, and then calculate similarity from the regression results.

What if, instead, we could construct a measure of artistic similarity directly based on the raw text descriptions? This would allow us to conveniently circumvent hedonic regressions and directly evaluate similarity across art pieces, since most hedonic information is probably already captured in these textual descriptions. To quantify similarity in this manner, we look to computer science. We can represent each observation, consisting of unique words, as a vector of words, where is the total number of unique words in our entire dataset. Many papers exist on how to calculate similarity between these word vectors,