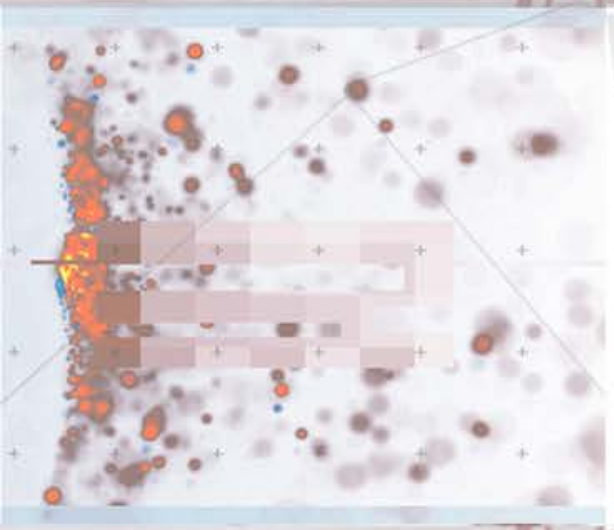

4.3. Pattern Evaluation: Null Invariance Measures In-Lecture Question

Question

DBLP is a computer science bibliography database. Suppose we want to study the correlation between authors Jure Leskovec (JL) and Hector Garcia-Molina (HG) based on their co-authorship, which is shown in the table.

Which of the following measures might be appropriate? Check all that apply.

- ❑ Lift
- ❑ χ^2
- ❑ Cosine
- ❑ Kulcyzynski
- ❑ AllConf

	JL	¬JL
HG	1	430
¬HG	133	2.7e6

Question

DBLP is a computer science bibliography database. Suppose we want to study the correlation between authors Jure Leskovec (JL) and Hector Garcia-Molina (HG) based on their co-authorship, which is shown in the table.

Which of the following measures might be appropriate? Check all that apply.

- Lift
- χ^2
- Cosine
- Kulcyzynski
- AllConf

	JL	¬JL
HG	1	430
¬HG	133	2.7e6

- **Answers:** Cosine, Kulcyzynski, AllConf
- **Explanation:** Since both JL and HG have published hundreds of papers and they have only co-authored one paper, JL and HG are negatively correlated. We have Cosine(JL, HG) = 4.1e-3, Kulc(JL, HG) = 4.9e-3, and Allconf(JL, HG) = 2.3e-3, which are small and agree with our observation.
- However, we have Lift(JL, HG) = 46.87 and χ^2(JL, HG) = 44.79, which are large due to a large number of null transactions. Therefore, Lift and χ^2 are not appropriate in this case and we should use measures that are null-invariant .