Provenance-Only Integration

Ashish Gehani    Dawood Tariq

SRI
Integration Challenges

- Metadata variation:
  - Abstraction levels
  - Completeness
  - Identifiers
  - Semantics

- Querying requires:
  - Record assembly
  - Reconciling syntax
  - Mapping semantics
Related Work

- Data integration

- Provenance integration
  - Semantic web (Umuhoza 2012)
  - Grid computing (Zhao 2008)
  - System interoperability (Angelino 2011)
  - Cross-organization sharing (Allen 2011)
Provenance-Only Integration

- Single underlying activity
- Multiple views of it
- Partial overlap in metadata
Speech Processing Hot Spots
Basic Provenance-Only Integration

- Provenance from two vantage points
- Need to integrate the two

Approach:
- Define matching threshold $\tau$
- Merge vertex pair if $\tau$-similar
- Merge edge pair if $\tau$-similar
- Cost from conflating owners

Goal:
- Minimize $\tau$
- Keep cost < tolerance $\Upsilon$
Android Provenance

- Security analysis
  - System-wide monitoring
- Resource-constrained
  - Disrupts power management
  - Blinded by garbage collection
- Multiple abstraction levels
  - Kernel interface
  - Inter-application (Binder)
  - Provenance-only integration
Fast Integration

- Integrate all $\tau$-similar elements
- Don’t have to find matching pairs
  - Avoids subgraph isomorphism problem
- Separate vertex, edge matching thresholds
- Thresholds are input now
- Cost is per match now

Approach:
- Merge $\tau_v$-similar vertices, if cost $< \gamma$
- Merge $\tau_e$-similar edges, if cost $< \gamma$
False Integration → High Cost

![Graph showing the relationship between Cost (τ) and Threshold (τ)]

- Cost (τ) decreases as Threshold (τ) increases.
Integration as Abstraction

Graph size (vertices) vs. Threshold (τ)
Fidelity of Attribution

![Graph showing the process count (vertices) vs. threshold (τ) for different values of γ: γ = 0, γ = 1, γ = 2, and γ = ∞. The graph illustrates how the process count decreases as the threshold increases, with each line representing a different value of γ.](image-url)
Conclusion

- Provenance-only integration
- Basic form as constrained optimization
- Fast version → automated abstraction

Acknowledgement
- TaPP ’14 organizers, reviewers
- US NSF Grant IIS-1116414

URL: http://data-provenance.googlecode.com

Email: ashish.gehani@sri.com

Questions?