On the Topological Landscape of Web Services Matchmaking

Hyunyoung Kil, Seog-Chan Oh, and Dongwon Lee
The Pennsylvania State University
USA

Motivation

- How useful are current web services?
  - Very little research
- Web services in WSDL form a network
  - Network analysis technique
- The topology of networks often help understand its behavior:
  - How to describe a large network quantitatively?
  - How were networks formed?
  - What are the consequences of a specific network organization?
WSDL

- Web Services Definition Language
- WSDL provides a way for service providers to describe the basic format of web service requests over different protocols or encodings
- WSDL is used to describe what a web service can do, where it resides, and how to invoke it
- Similar to IDL of CORBA
WSDL

- A **web service**, ws, consists of operations
- An **operation**, op, consists of input and output parameters: \( op(IN, OUT) \)
  - One-way
  - Request-response
  - Solicit-response
  - Notification
- A **parameter**, \( p \), has name and type
  - \( p(name, type) \)

---

Example

```xml
<message name='findRestaurant_Request'>
  <part name='zip' type='xs:string'>
  <part name='foodPref' type='xs:string'>
</message>

<message name='findRestaurant_Response'>
  <part name='zip' type='xs:string'>
  <part name='phone' type='xs:integer'>
</message>

<portType name='allRestaurant'>
  <operation name='findRestaurant'>
    <input message='findRestaurant_Request'/>
    <output message='findRestaurant_Response'/>
  </operation>
</portType>
```

- **op**: findRestaurant
- **IN**: \{zip, foodPref\}
- **OUT**: \{zip, phone\}
Matchmaking Framework: Parameter Matching

- For two parameters:
  - $p_1$(name$_1$, type$_1$) and $p_2$(name$_2$, type$_2$),
  - **Type-Match**: $p_1$ type-matches $p_2$ iff
    1. $p_1$.type == $p_2$.type or
    2. $p_1$.type is derived from $p_2$.type.
  - **Name-Match**: $p_1$ name-matches $p_2$ iff
    $\text{dist}(p_1$.name, $p_2$.name) $\leq$ threshold
  - $p_1$ matches $p_2$, ($p_1 \sim p_2$), if $p_1$ and $p_2$ are type-match and name-match
  - Eg. $p$(“password”, xs:string) $\sim$ $q$ (“pwd”, xs:string)

Matchmaking Framework: Operation Invocation

- For two operations
  - $\text{op}_1$(IN$_1$, OUT$_1$) and $\text{op}_2$ (IN$_2$, OUT$_2$),
  - **Full Invocation (FI)**
    - $\text{op}_1$ fully invokes $\text{op}_2$ if for every mandatory input parameter $p$ in IN$_2$, there exists an output parameter $q$ in OUT$_1$ such that $q \sim p$.
  - **Partial Invocation (PI)**
    - $\text{op}_1$ partially Invokes $\text{op}_2$ if there exists any mandatory input parameter $p$ in IN$_2$, an output parameter $q$ in OUT$_1$ such that $q \sim p$. 
Matchmaking Framework: Web Service Network Model

WS₁

OP₁

INPUT

OUTPUT

WS₂

OP₂

INPUT

OUTPUT

Matchmaking Framework: Web Service Network Model

WS₁

OP₁

INPUT

OUTPUT

WS₂

OP₂

INPUT

OUTPUT

~
Matchmaking Framework: Web Service Network Model

ws: $WS_1 \rightarrow WS_2$

op: $OP_1 \rightarrow OP_2$

p: $p_1 \rightarrow p_2$

Matchmaking Framework: Web Service Network Model

ws: $WS_1 \rightarrow WS_2$

op: $OP_1 \rightarrow OP_2$

p: $p_1 \rightarrow p_2$

p: $p_3 \rightarrow p_4$

p: $p_5 \rightarrow p_6 \rightarrow p_7$

p: $p_5 \rightarrow p_6 \rightarrow p_7$
Matchmaking Framework: Web Service Network Model

WS₁

<table>
<thead>
<tr>
<th>INPUT</th>
<th>OP₁</th>
<th>p₁</th>
<th>p₂</th>
<th>p₃</th>
<th>p₄</th>
</tr>
</thead>
</table>

WS₂

<table>
<thead>
<tr>
<th>OUTPUT</th>
<th>OP₂</th>
<th>p₅</th>
<th>p₆</th>
<th>p₇</th>
</tr>
</thead>
</table>

INPUT={p₄, p₈}

Partial Invocation

op: OP₁ → OP₂

Full Invocation

op: OP₁ → OP₂

Plan of Study

- Use real-world public WSDL files
- Study the topology of various networks formed from the downloaded WSDL files
  - Small-world
  - Power-law
- Use different distance metrics and thresholds in parameter matching
  - Exact, Cosine with TF-IDF, WordNet
Data Pre-Processing

- Data Gathering: 2,100 WSDLs
  - 1,554 files from Fan et al. + top-1,000 WSDLs from Google
- WSDL Validation: 1,360 WSDLs are left
  - After removing 740 invalid WSDL files based on WSDL standard.
- WSDL De-duplication: 984 WSDLs are left
  - After removing 376 duplicate WSDL files at operation level.
- Type Flattening
  - When a parameter has not a simple type, we flatten the type.
  - ex) p1(address, addressType1), where addressType1 is (integer zipcode, string street, string city, string state), changes to a set of parameters, p11(zipcode, integer), p12(street, string), p13(city, string) and p14(state, string).
- Data Cleaning
  - Improve data quality, e.g., replacing too general names such as “return”, “result”.

Small-World Network Model

- Watts & Strogatz (1998)
- Arrange N nodes in a ring and connect each node to k others in each direction. (each node degree = 2k.)
- With probability p “re-wire” each connection from node i to a new node
- Small-world networks show both regularity and randomness
  - Highly clustered & small shortest distance
Small-World Network Model

- $L$: average shortest distance among all reachable pairs of nodes
- $C$: average clustering coefficient of all nodes

- $\text{Index}^{\text{SN}} = |C_{\text{actual}} - C_{\text{random}}| / |L_{\text{actual}} - L_{\text{random}}|
- If a network is small-world, then
  - $C_{\text{actual}} >> C_{\text{random}}$
  - $L_{\text{actual}} \sim L_{\text{random}}$
  - $\Rightarrow$ Large $\text{Index}^{\text{SN}}$

Result 1: Small World

<table>
<thead>
<tr>
<th>Matching Scheme</th>
<th>Network</th>
<th>$L_{\text{actual}}$</th>
<th>$C_{\text{actual}}$</th>
<th>$L_{\text{random}}$</th>
<th>$C_{\text{random}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact matching</td>
<td>Ms</td>
<td>1.3420</td>
<td>0.2418</td>
<td>2.2298</td>
<td>0.3143</td>
</tr>
<tr>
<td></td>
<td>Ms</td>
<td>1.3920</td>
<td>0.2156</td>
<td>2.2298</td>
<td>0.3143</td>
</tr>
<tr>
<td></td>
<td>Ms</td>
<td>2.0304</td>
<td>0.2218</td>
<td>2.2298</td>
<td>0.3143</td>
</tr>
<tr>
<td></td>
<td>Ms</td>
<td>2.1704</td>
<td>0.2218</td>
<td>2.2298</td>
<td>0.3143</td>
</tr>
</tbody>
</table>

| Cosine (0.96)  | Ms      | 1.9000              | 0.3146              | 2.2298              | 0.3143              |
|                | Ms      | 2.0304              | 0.2218              | 2.2298              | 0.3143              |
|                | Ms      | 2.1704              | 0.2218              | 2.2298              | 0.3143              |
|                | Ms      | 2.3004              | 0.2218              | 2.2298              | 0.3143              |

| Cosine (0.75)  | Ms      | 1.9000              | 0.3146              | 2.2298              | 0.3143              |
|                | Ms      | 2.0304              | 0.2218              | 2.2298              | 0.3143              |
|                | Ms      | 2.1704              | 0.2218              | 2.2298              | 0.3143              |
|                | Ms      | 2.3004              | 0.2218              | 2.2298              | 0.3143              |

| WordNet (0.96)| Ms      | 2.8900              | 0.2838              | 2.2298              | 0.3143              |
|                | Ms      | 2.9400              | 0.2838              | 2.2298              | 0.3143              |
|                | Ms      | 2.9900              | 0.2838              | 2.2298              | 0.3143              |
|                | Ms      | 3.0400              | 0.2838              | 2.2298              | 0.3143              |

| WordNet (0.75)| Ms      | 2.8900              | 0.2838              | 2.2298              | 0.3143              |
|                | Ms      | 2.9400              | 0.2838              | 2.2298              | 0.3143              |
|                | Ms      | 2.9900              | 0.2838              | 2.2298              | 0.3143              |
|                | Ms      | 3.0400              | 0.2838              | 2.2298              | 0.3143              |
Scale-Free Network Model

- Barabasi & Albert (1999)
- Small number of nodes with many links (Hubs) and many nodes with only a few links
- Scale-free link distribution often follows power-law
  - the proportion of nodes with a given number of links $n$ is $P(n) = 1/n^k$.
- Network grows by addition of new nodes with preferential attachment to the existing nodes based on their number of links with a probability proportional to their degrees

Result 2: Power-law

- Complexity of Web services
Result 2: Popularity of Parameter Names

Result 2: Out-degree Distribution
Conclusion

- Using five matching schemes and three network granularities:
  - Public web services networks show small-world network property
  - Public web service networks follow power-law like distribution pattern
- Semantic Web is needed:
  - Using only exact matching in web service composition is too rigid
  - Using approximate matching helps but not sufficient
  - Semantic web services matchmaking is needed (eg, WSDL-S)
- More study is needed:
  - Giant component size, diameter, formation of networks, etc.