Title: Sparta: Practical Anonymity with Long-Term Resistance to Traffic Analysis Abstract: Existing metadata-private messaging systems are either non-scalable or vulnerable to long-term traffic analysis. Approaches that mitigate traffic analysis attacks often suffer from unrealistic and unimplementable assumptions or impose system-wide bandwidth restrictions, degrading usability, and performance. In this work, we present a new model for metadata-private communication systems---deferred retrieval---that guarantees traffic analysis resistance under realistic, implementable user assumptions. We introduce Sparta systems, practical and scalable instantiations of deferred retrieval that are distributable, achieve high throughput, and support multiple concurrent conversations without message loss. Specifically, we present three Sparta constructions optimized for different scenarios: (i) low-latency, (ii) high-throughput in shared-memory environments (multi-thread implementations), and (iii) high throughput in shared-nothing (distributed) environments. Our low latency Sparta supports latencies of less than one millisecond, while our high-throughput Sparta can scale to deliver over 700,000 100B messages per second on a single 48-core server.