Copula Analysis for Statistical Network Calculus

Kui Wu, Fang Dong, Venkatesh Srinivasan

Computer Science Department
University of Victoria
Canada, B.C.

wkui@uvic.ca

Overview

Background in finance domain

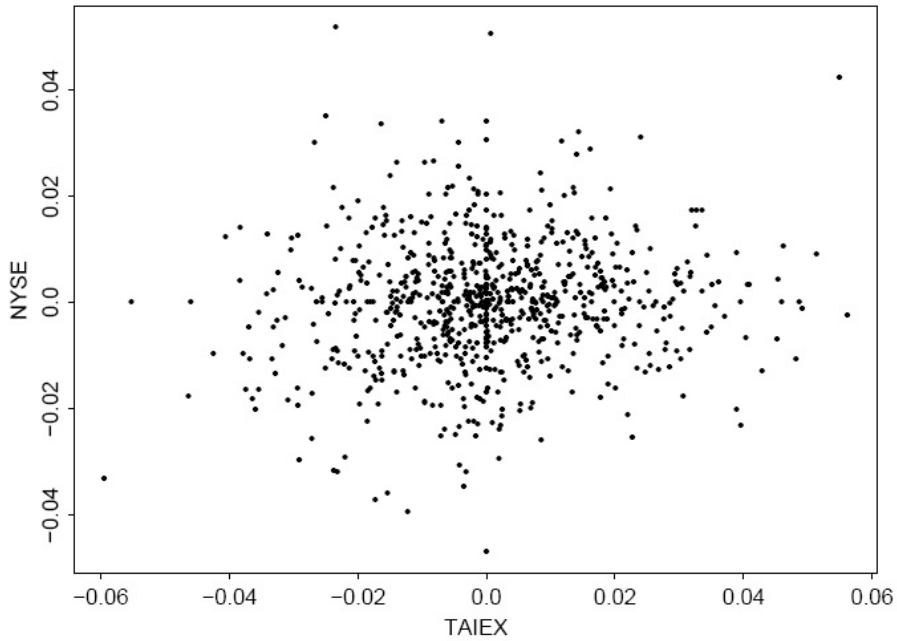


Figure : Scatter plot of the daily log returns of the New York Stock Exchange composite price index (NYSE) versus Taiwan weighted stock index (TAIEX) from 2001 to 2003. (The figure is cited from [**chiou2008copula**])

[chiou2008copula] Chiou, Shang C., and Ruey S. Tsay. "A copula-based approach to option pricing and risk assessment." Journal of Data Science 6.3 (2008): 273-301.

What's the copula?

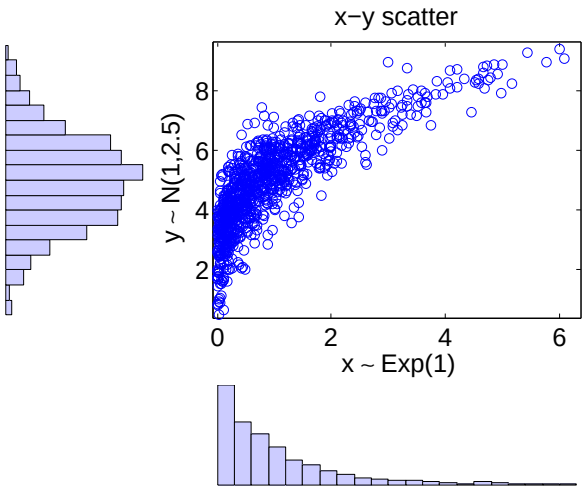


Figure : Each point in the above figure is a sample pair (x', y')

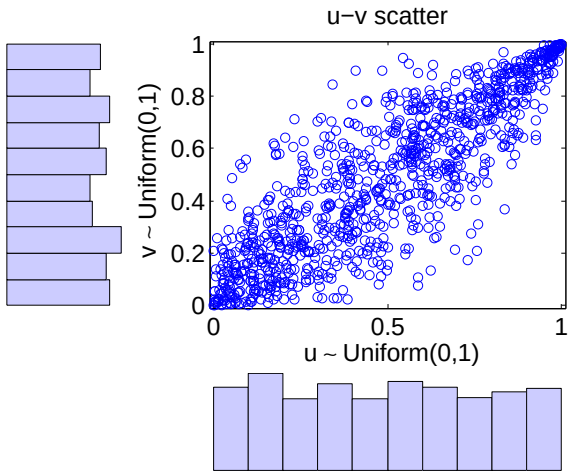


Figure : Each point in the above figure is a sample pair (u', v')

(x', y') is mapping to (u', v') by marginal distribution function, *i.e.*,

$$u' = F_x(x') = Prob(x \leq x')$$
$$v' = F_y(y') = Prob(y \leq y')$$

Copula $C(u', v') = Prob(u \leq u', v \leq v')$

Definition of Copulas

Definition (Copula)

Copula is a multivariate distribution function which must satisfy that the marginal distribution of each argument is a uniform distribution on [0,1].

Theorem (Sklar's theorem)

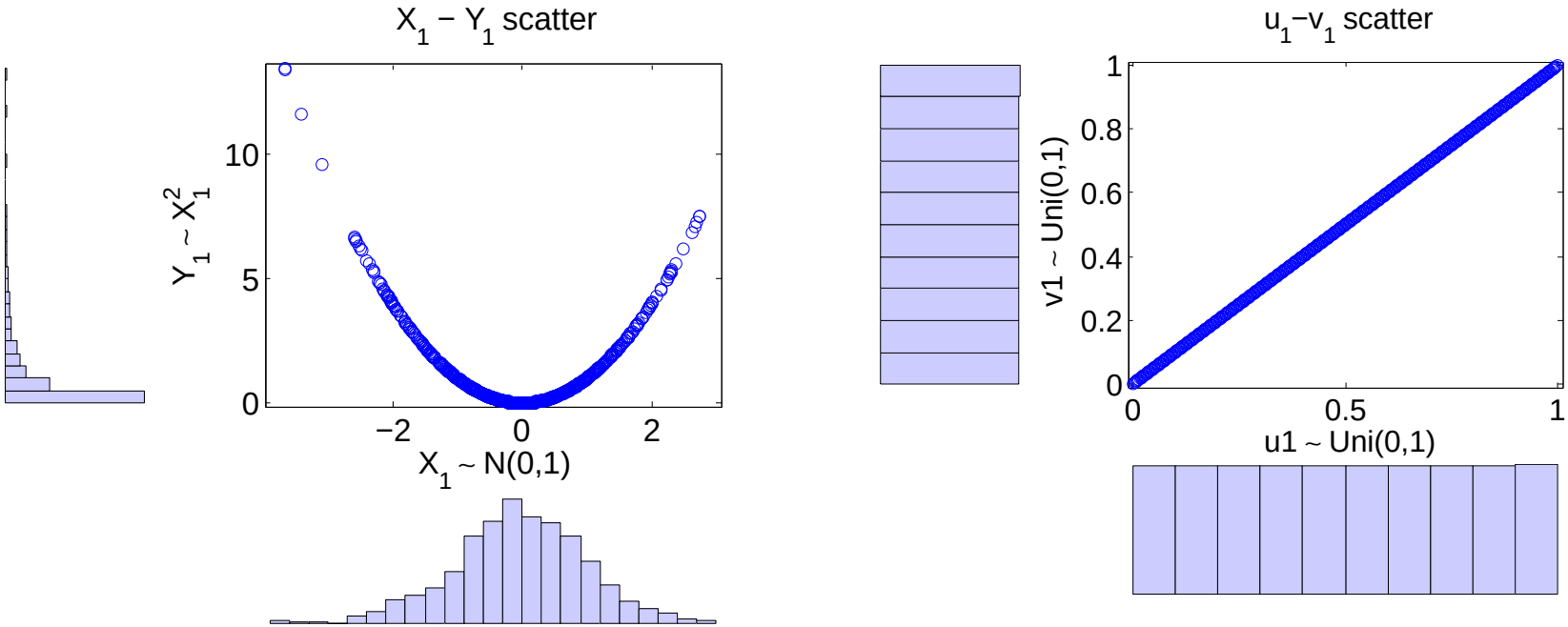
Let F be an N-dimensional joint distribution function with continuous margins $F_1, F_2, \ldots, F_N$. Then F has a unique copula representation:

$$F(x_1, x_2, \ldots, x_N) = C(F_1(x_1), F_2(x_2), \ldots, F_N(x_N)).$$

- Copula is a function that links univariate marginals to their joint distribution.
- Copula is independent of both marginals and joint distribution. Several popular families have been proposed.
- Given marginals, joint distribution is computable by copula modelling.

Why Are Copulas Powerful?

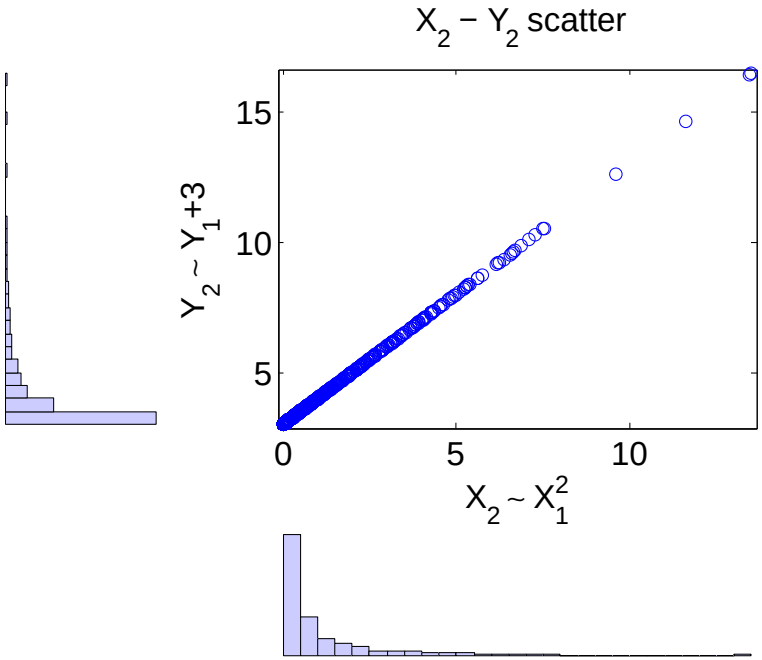
$Y_1 = X_1^2$

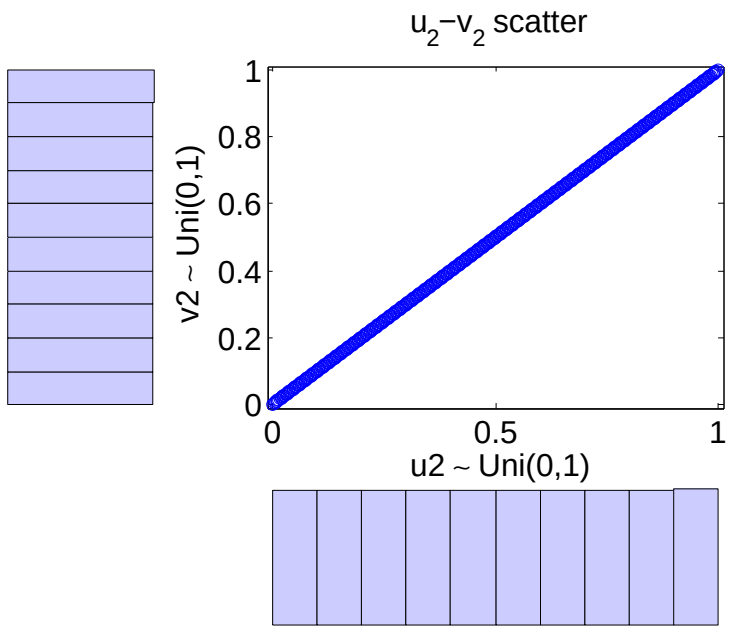


- Pearson Correlation $\rho(X_1, Y_1) = \frac{cov(X_1, Y_1)}{\sigma_{X_1}\sigma_{Y_1}} = 0$
- Kendall's Tau $\rho_\tau(X_1, Y_1) = 4\int_0^1\int_0^1 C(u_1, v_1)dC(u_1, v_1) - 1 = 1$
- **Copula measures functional dependence, while the Pearson correlation only measures linear dependence.**

Why Are Copulas Powerful? (Cont.)

$X_2 = X_1^2, Y2 = Y1 + 3$



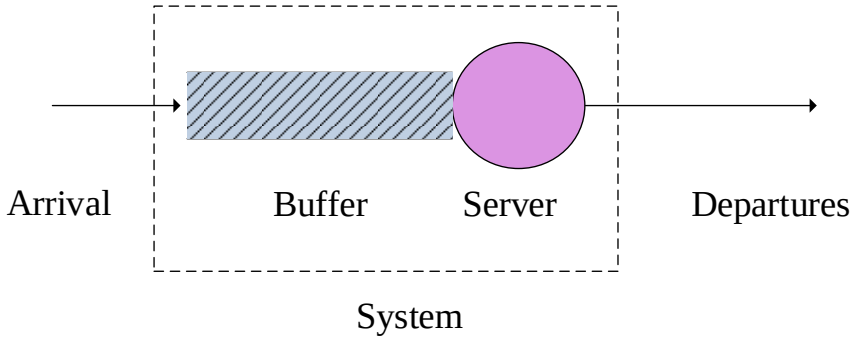


$$\rho(X_2, Y_2) = 1, \qquad \rho_\tau(X_2, Y_2) = 1$$
$$\rho(X_1, Y_1) = 0, \qquad \rho_\tau(X_1, Y_1) = 1$$

Theorem (The invariant property of copulas)

Let X and Y be continuous random variables with copula C_{XY}. C_{XY} is invariant under strictly increasing transformations of X and Y.

Statistical Network Calculus (StatNC)



$$\text{Arrival Curve} \quad Prob\{\sup_{0\le s\le t}\{A(s,t)-\alpha(t-s)\}>x\}\le f(x)$$
$$\text{Service Curve} \quad Prob\{A\otimes\beta(t)-A^*(t)>x\}\le g(x)$$
$$\Downarrow$$
$$\text{Backlog} \quad Prob\{\mathcal{B}(t)>x\}\le (f\otimes g)(x-\alpha\oslash\beta(0))$$
$$\text{Delay} \quad Prob\{\mathcal{D}(t)>h(\alpha+x,\beta)\}\le (f\otimes g)(x)$$

The values of the bounding function on the right hand of the inequations are expected to be as small (tight) as possible.

Statistical Network Calculus (StatNC)

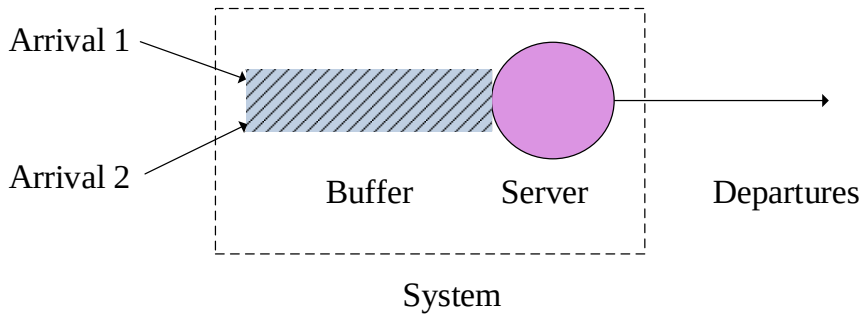
Arrival Curve $\quad Prob\{\sup_{0\leq s\leq t}\{A(s,t)-\alpha(t-s)\}>x\}\leq f(x)$

Statistical Method

- Characterize the arrival traffic by random variable
  let $X = \sup_{0\leq s\leq t}\{A(s,t)-\alpha(t-s)\}$ as the **statistic** of arrival A.
  Sampling it along time series, it can be regarded as random variable.
- The arrival curve of traffic flow turns to be the complementary distribution of the random variable: $Prob\{X>x\}=\bar{F}_X(x)\leq f(x)$.

Integration of Copulas into StatNC

Two Flows Situation



Goal: to capture dependence between A_1 and A_2, and characterize aggregate traffic $A_1 + A2$.

Statistical Method: the goal turns to capture dependence between X_1 and X_2, and determine the complementary distribution of the sum $Z = X_1 + X_2$

Integration of Copulas into StatNC (cont.)

Problem Formation: If we know two flow curves $\bar{F}_{X_1}$ and $\bar{F}_{X_2}$, how to calculate the aggregate flow curve $\bar{F}_Z$, where $Z = X_1 + X_2$?

Traditional Solutions:

- General Case (No matter independent or not)

$$\bar{F}_Z(z) \leq (\bar{F}_{X_1} \otimes \bar{F}_{X_2})(z), \otimes \text{is (min,+) convolution.}$$

- Independent Case

$$\bar{F}_Z(z) = 1 - (\bar{F}_{X_1} * \bar{F}_{X_2})(z), * \text{is the Stieltjes convolution operation.}$$

New Solution with the Copula:

$$\bar{F}_Z(z) = 1 - \iint_{x+y<z} \mathrm{d}C(F_{X_1}(x), F_{X_2}(y))$$

Integration of Copulas into StatNC (Cont.)

Copula case: Let Z be the sum of two random variables X and Y. Then

$$\hat{\bar{F}}_Z(z) \geq \bar{F}_Z(z) \geq \check{\bar{F}}_Z(z), \tag{1}$$

where

$$\hat{\bar{F}}_Z(z) = 1 - \sup_{x+y=z} \{W(F_X(x), F_Y(y))\}, \tag{2}$$

$$\check{\bar{F}}_Z(z) = 1 - \inf_{x+y=z} \{\bar{W}(F_X(x), F_Y(y))\}, \tag{3}$$

$$W(u, v) = [u + v - 1]^+, \tag{4}$$

$$\bar{W}(u, v) = [u + v]_1. \tag{5}$$

Integration of Copulas into StatNC (Cont.)

Numerical Examples: $X_1 \sim Exp(0.5), X_2 \sim Exp(0.5), Z = X_1 + X_2$

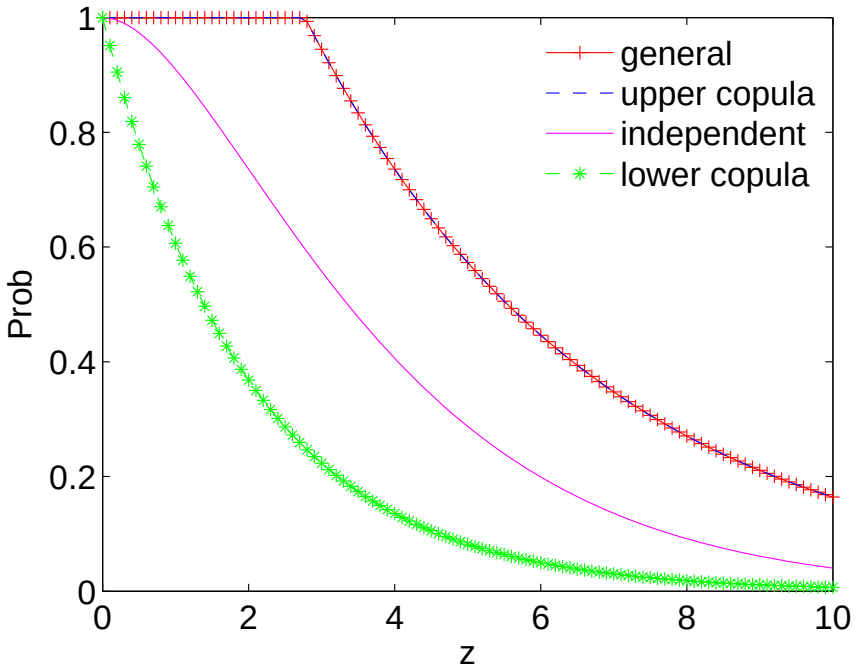
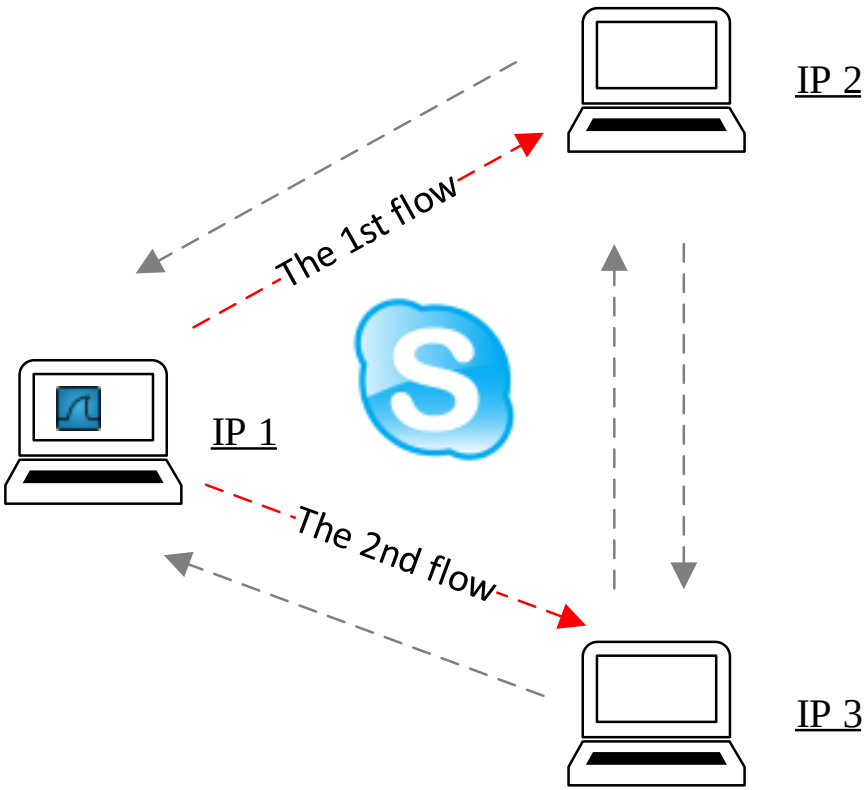


Figure : Arrival Curve of Aggregate Flow $Z = X_1 + X_2$

Promising aspects of copulas:

- Copula could measure the dependence between two traffic flows more accurately
- Copula-based analysis for aggregate traffic leads to tighter bounds than general bounds
- The upper and lower bounds derived by copulas show the range that statistical network calculus can achieve.

Experiment Setting



- A Skype group call with three clients
- The traffic data of two outflows from IP 1 are captured by Wireshark.
- There are three independent experiments, each records traffic for more than 20 minutes. The corresponding traffic data are stored in three datasets, Dataset 1, Dataset 2 and Dataset 3.

Traffic Modelling

We define a random variable $\mathbf{a}$ to represent traffic sent per second. The observed samples of this random variables are denoted as $\hat{a}$,

$$A(s,t) = \sum_{i=s+1}^{t} \hat{a}^i,$$

where $\hat{a}^i$ is the observed value of $\mathbf{a}$ in ith second.

- We model traffic with a for the sake of the preservation of characteristics of real-world traffic.
- The goals for modelling: 1) study the marginals of $\mathbf{a}_1$ and $\mathbf{a}_2$; 2) study the dependence structure between $\mathbf{a}_1$ and $\mathbf{a}_2$.

A Study on Real-world Traffic

Marginal Study of $\mathbf{a}_1$ and $\mathbf{a}_2$

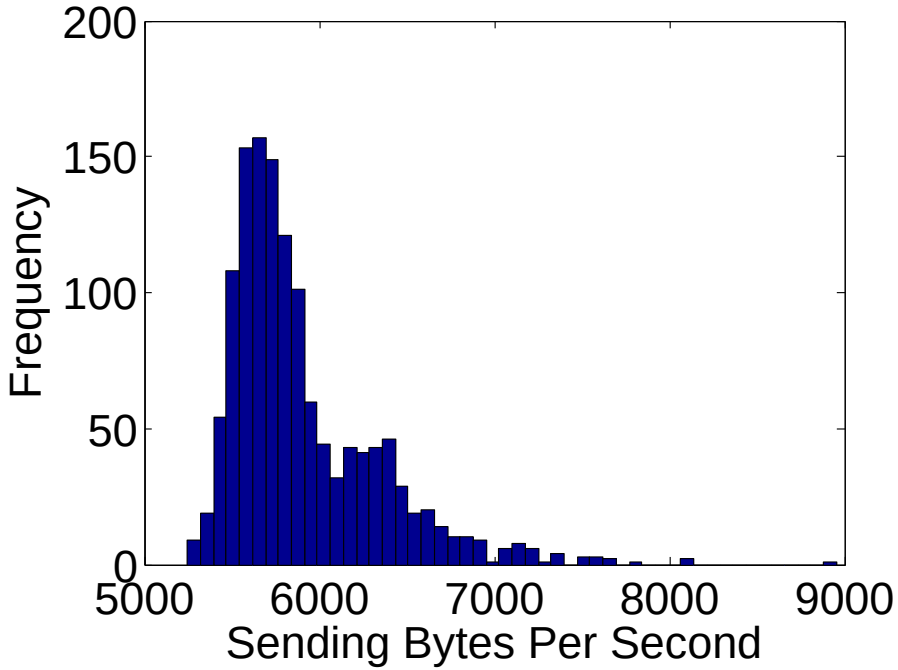


Figure : Histogram of samples of $\mathbf{a}_1$ in Dataset 1

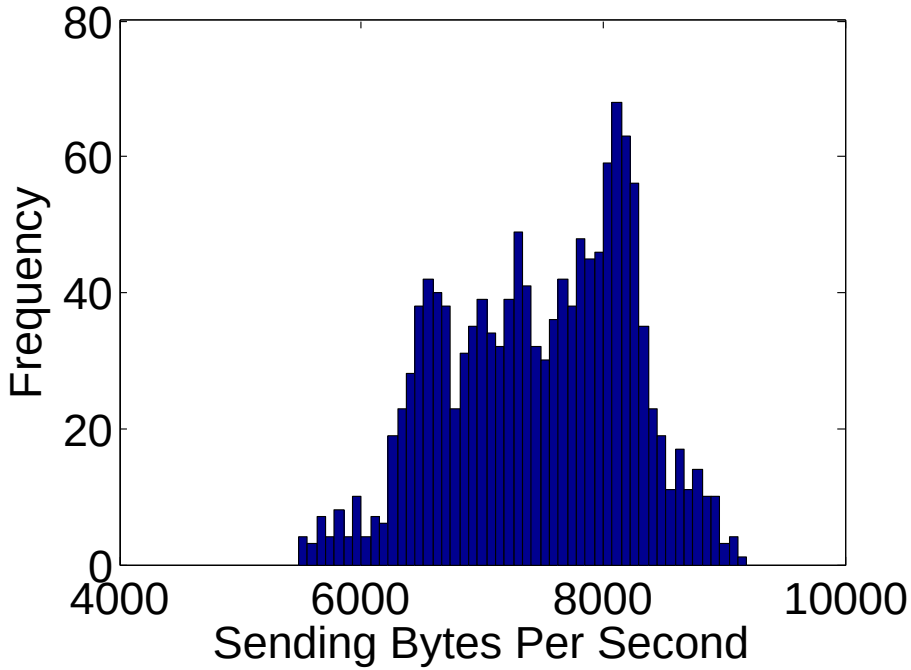


Figure : Histogram of samples of $\mathbf{a}_2$ in Dataset 1

Passed the Kolmogorov-Smirnov test that a_i conforms to the mixed normal distribution with distribution function:

$$F(x) = \pi\Phi(\frac{x-\mu_1}{\sigma_1}) + (1-\pi)\Phi(\frac{x-\mu_2}{\sigma_2}).$$

A Study on Real-world Traffic

Copula Study Between $\mathbf{a}_1$ and $\mathbf{a}_2$

We choose three of Archimedean parametric (θ) copulas (Clayton, Frank, Gumbel) to model the copula between $\mathbf{a}_1$ and $\mathbf{a}_2$.

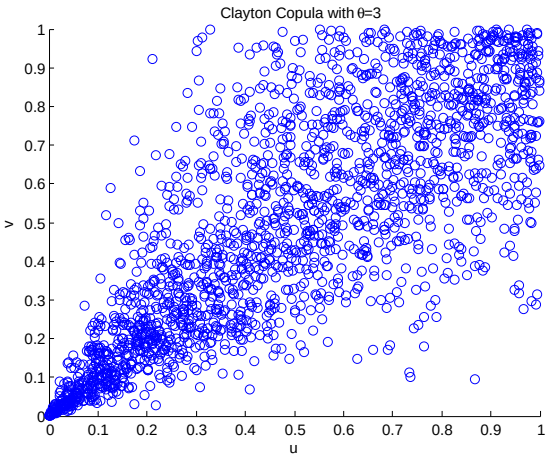


Figure : Clayton Copula with $\theta = 3$

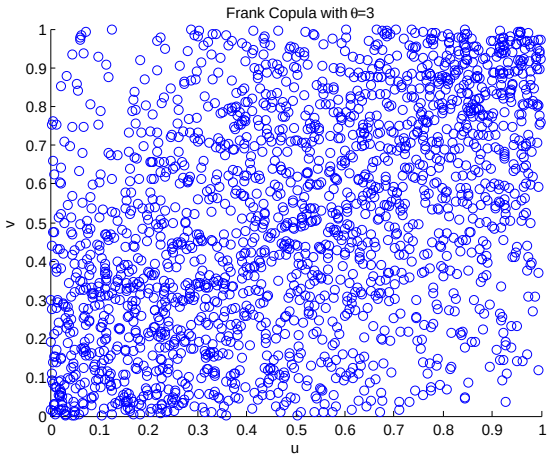


Figure : Frank Copula with $\theta = 3$

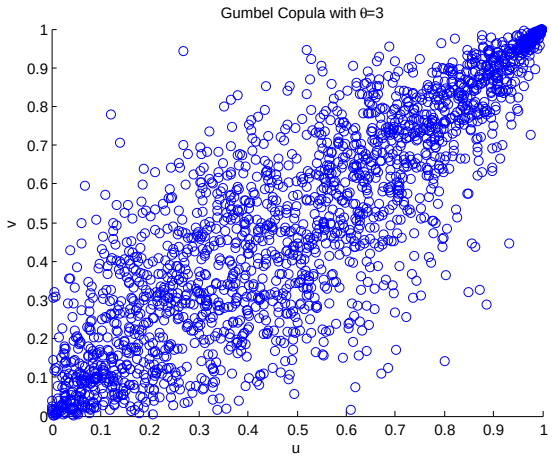


Figure : Gumbel Copula with $\theta = 3$

- Parameter θ is related directly to the Kendall's tau.
- The three copulas has three extreme features on tail dependence (Tail dependence is also copula based dependence).

Copula Study Between $\mathbf{a}_1$ and $\mathbf{a}_2$

		Dataset 1	Dataset 2	Dataset 3
Gumbel	θ	1.1464	1.0597	1.6791
	P-value	0.41	0.5	0.94
Frank	θ	1.2531	0.4483	4.465
	P-value	0.23	0.41	0.4
Clayton	θ	0.2057	0.0327	0.7574
	P-value	0.07	0.21	0

Table : 'Blanket' goodness of fit test for copula between $\mathbf{a}_1$ and $\mathbf{a}_2$ across three datasets.

Remarks:

1. P-Value is a measure of fit, with larger values being better.
2. Dependence between flows is best characterized by **Gumbel copula**.

Performance analysis

Definition (Backlog for Constant Service Rate System)

Given the arrival $A(s,t)$ and a constant service rate R, the backlog $\mathcal{B}(t)$ is:

$$\mathcal{B}(t) = \sup_{0 \le s \le t} \{A(s,t) - R(t-s)\}.$$

Recall that statistic of A, $X = \sup_{0 \le s \le t}\{A(s,t) - \alpha(t-s)\}$

- In the special case that the service rate is constant, **the backlog turns out to have the same form with the statistic of** A. Statistical method is applied on backlog directly. The backlog bounds from statistical method will be consistent with the derived bounds.
- We define a random variable $\mathbf{B}$ to represents the backlog. The backlog bound now is the complementary distribution function of the sum of r.v.s for two subflows. $Prob\{\mathbf{B} > x\} = Prob\{\mathbf{B}_1 + \mathbf{B}_2 > x\}$

Copula Analysis with Simulated Traffic

Statistical Distribution of Backlogs

Null hypothesis: the random variable $\mathbf{B}_i$ $(i = 1, 2)$ conforms to the mixture of two normal distribution with parameters given by the parameter estimates. **(Can't be rejected based on the K-S test.)**

Random variable		$\mathbf{B}1$	$\mathbf{B}2$
Estimate of paramters	π	0.316657	0.31119
	$\mu 1$	6402.2	6912.625
	$\mu 2$	10741.37	12382.1
	$\sigma 1$	1650.444	2439.608
	$\sigma 2$	1930.165	4116.222
Statistical value D		0.021	0.0233
Degree of freedom		1000	
Critical values $D_{0.01}$		0.0515	

Table : Kolmogorov-Smirnov test for backlogs based on simulated dataset.

Remark: Statistical property of backlog is essentially inherited from that of simulated arrival traffic flows.

Backlog Bound Curves

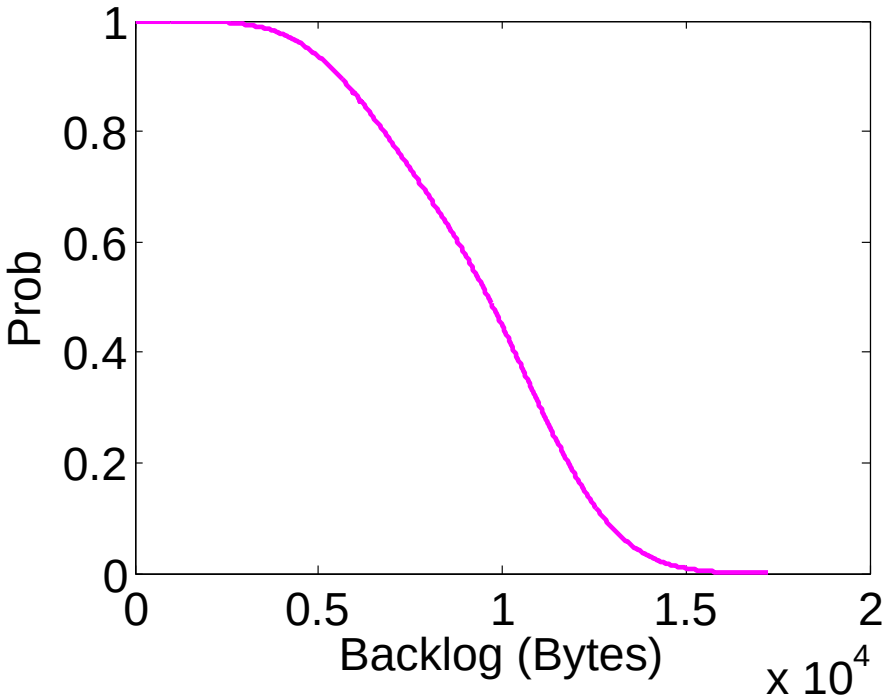


Figure : Backlog bound curve of flow A_1.

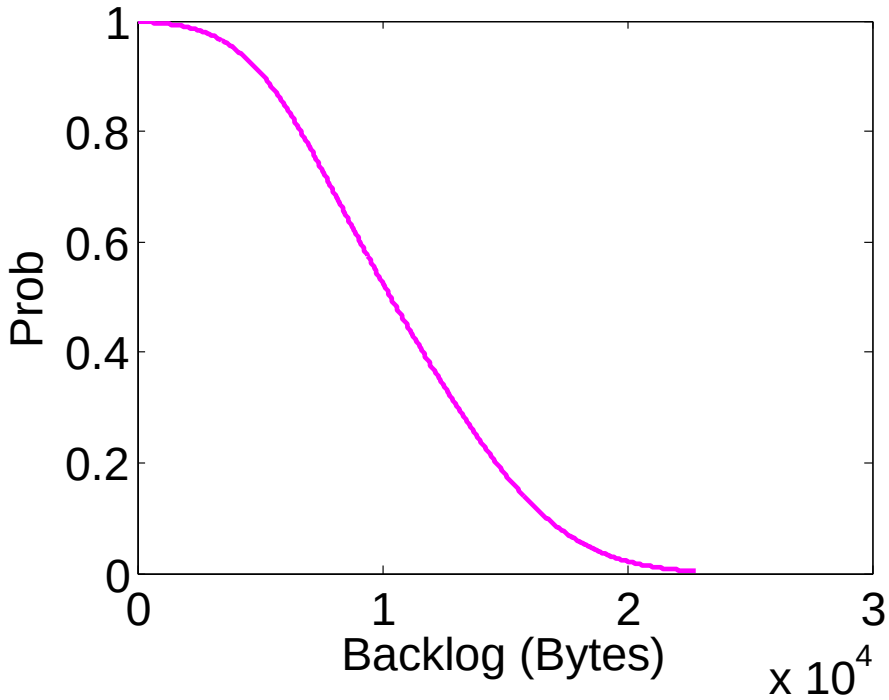


Figure : Backlog bound curve of flow A_2.

Copula-based Dependence Between backlogs

Gumbel	θ	2.48
	P -value	0.03
Frank	θ	9.3526
	P -value	0.21
Clayton	θ	3.5
	P -value	0.68

Backlog Bound for Aggregate Flow

General bound $Prob\{B(t) > x\} \leq \bar{F}_{\mathbf{B}_1} \otimes \bar{F}_{\mathbf{B}_2}$.

Backlog bound $Prob\{B(t) > x\} \leq 1 - \int\int_{b_1+b_2<x} dC(F_{\mathbf{B}_1}, F_{\mathbf{B}_2})$

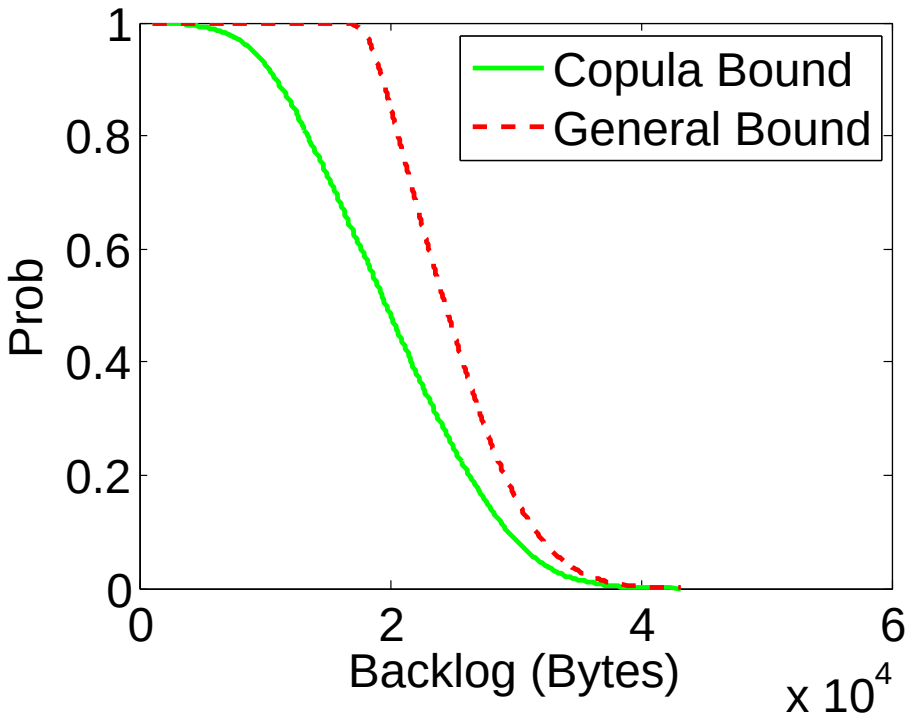


Figure : Backlog bound for aggregate traffic flow $A = A_1 + A_2$.

Conclusion and Future Work

1. Copula can be used as a statistical tool to capture network traffic dependence.
2. Copula analysis discloses the range that stochastic network calculus can achieve.
3. We show how to integrate copula analysis into the StatNC framework to provide tighter performance bounds.
4. A real-world case study as well as simulation evaluation demonstrate the copula analysis.
5. So far, we only study the **contemporaneous** dependence and its application in StatNC. It is more important to investigate the **temporal** dependence in the future.

Questions?

The End